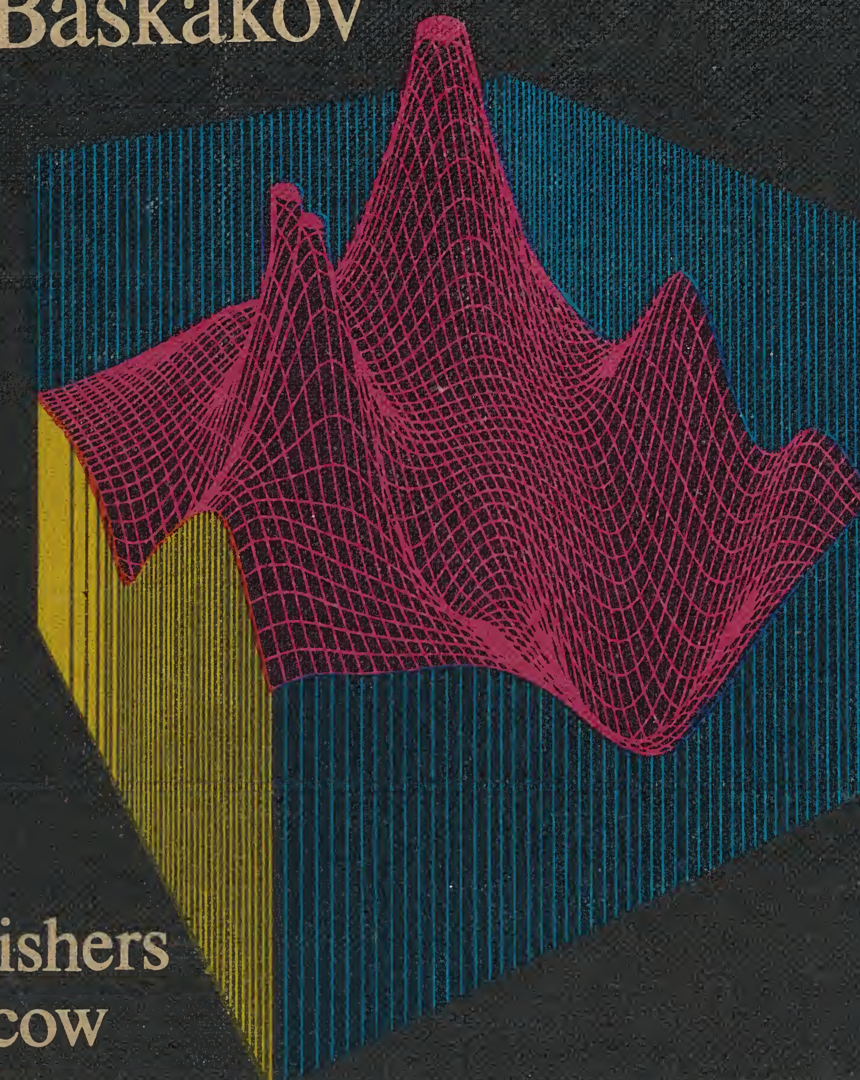
Signals and Circuits

S.I.Baskakov



Mir
Publishers
Moscow

S.I.Baskakov Signals and Circuits

С. И. Баскаков
Радиотехнические цепи
и сигналы
«Высшая школа» Москва

Signals and Circuits

S. I. Baskakov

Translated from the Russian
by
Boris V. Kuznetsov

Mir Publishers Moscow

First published 1986
Revised from the 1983 Russian edition

TO THE READER

Mir Publishers welcome your comments on the content, translation, and design of the book.
We would also be pleased to receive any suggestions you care to make about our future publications.
Our address is:
USSR, 129820,
Moscow, I-110, GSP,
Pervy Rizhsky Pereulok, 2.
Mir Publishers

На английском языке

Printed in the Union of Soviet Socialist Republics

Contents

Preface

The present book is a course on signals and circuits as it is taught in the USSR. This subject figures prominently among the fundamental disciplines essential to the expertise of communication engineers. Keeping pace with overall progress in science and technology and reflecting the current trends in component design and theory, this course combines and sets forth in a systematic way the most important principles in the field of communications.

In his work on the text, the author has been guided by the idea that material should be specifically tailored to the teaching practice at college. This approach has governed the selection of material and the degree of detail in its presentation: its pages contain what, as the author believes, the student can fully assimilate during the time allotted. Specific circuit types, their study and comparative analysis–all this belongs to the specialized subjects in communication engineering.

This text includes a wide variety of material and a wealth of concepts and techniques which will come the student's way for the first time. Ample space is devoted to mathematical tools of study. To link theory closer to practice, the chapters contain a great number of examples and problems giving the student deeper insight into the techniques of engineering analysis.

Subject-matter and structure. The book is in two parts. Part One, *Signals*, introduces the reader to the methods currently used to describe and study the properties of signals. Among other things, it covers the classification of signals, the fundamental principle of the geometric treatment of the signal space, the spectral and correlation analysis of deterministic signals, the theory of modulated signals, and the discrete presentation of band-limited continuous signals. The methods used to analyse and measure the characteristics of random signals are discussed in detail.

Part Two, *Circuits*, gives a systematic exposition of the principles underlying the analysis and determination of the response of linear and nonlinear circuits to both deterministic and random signals. Special emphasis is placed on the role of narrowband frequency-selective networks. Techniques are discussed that are used to synthesize linear two-terminal (one-port) and four-terminal (two-port) networks having a predetermined frequency response, and consideration is given to the passage of signals through linear parametric circuits. In the pages devoted to nonlinear lag-free (zero-memory) circuits, the reader will learn about the most important forms of signal transformations, such as modulation, detection, multiplication, and frequency conversion. Material is included on the theory of harmonic self-excited oscillators and the more recent trends that have come into being only recently in the wake of advances in microelectronics. These include active filters for analog signals and, as a most promising technique, digital signal

filtering. Finally, elements of the theory of optimal linear filtering as applied to deterministic and random signals are discussed.

When the reader opens the book, he will undoubtedly notice its makeup. In addition to the text in the body of a page, its margins display auxiliary, supplementary and graphical material. More specifically, the material on the margins includes.

1. REMINDERS referring to the previous courses, such as physics and circuit theory.
2. REFERENCE DATA, such as basic physical constants, tabulated integrals, and the like.
3. SHORT NOTES which have as their objective to draw the reader's attention to the relation existing among various subjects, the methods common to communication engineering and other, sometimes remote, fields of pure and applied science.
4. AUXILIARY DRAWINGS which will not, as a rule, be referred to in the text, but are an integral part of the study material. Apart from helping reduce the size of the book, they give it, as far as practicable, the flavour of a live lecture.
5. LABELS to help the reader organize his work on the book:

■ Here the text describes a principle of primary importance to communication theory and engineering.

● The reader is alerted to the fact that a new concept has been formulated and it must be memorized.

▲ The reader is advised to work a problem from among those given at the back of the respective chapter; the problem illustrates some point (or points) set forth in the body of the text.

Today the college student has to assimilate material at a high pace. That is why he must schedule his academic activity carefully. Wishing to help him in this matter, the author has taken special pains to grade the material offered for assimilation and recapitulation optimally. Each chapter corresponds to a major topic of the lecture course. The basic structural unit of a chapter is a section roughly corresponding to a complete lecture.

At the back of each chapter there is a SUMMARY. A firm knowledge of the summaries is mandatory. There are also REVIEW QUESTIONS which will be especially useful to the student in preparation for his examinations. The PROBLEMS section covers material for the student's work on his own and corresponds to the respective chapter and the degree of complexity. Each chapter includes ADVANCED PROBLEMS intended for students who have special interest in communication theory and tend towards research work.

The book is based on the lectures that the author has been reading at the Radio Engineering Department of the Moscow Power Institute. The author wishes to express his deep gratitude to his colleagues, especially professor G. D. Lobov and Docent V. P. Zhukov for their unfailing support and valuable suggestions.

Svyatoslav I. Baskakov

Introduction

Communication engineering is both a science and an art whose objectives are:

(1) to study the principles underlying the generation, amplification, emission, and reception of electromagnetic waves falling within the radio-frequency (r.f.) range;

(2) to put these waves to practical use for the purposes of transmitting, storing and converting information.

In its early days, after A.S. Popov invented radio in 1895, telecommunication was mainly based on the use of wavelengths of several hundred or even thousand metres. Today, the field of telecommunications has expanded enormously. Radio communications, television, radio control, radar, and radio navigation are only a few of the many divisions of telecommunications.

The science that has to do with the physical principles of radio engineering is known as *radio physics*. It is a division of the applied natural sciences, closely related to the fundamental fields such as quantum mechanics, solid-state physics, to name but a few.

Communication engineering in the USSR has been pushing in many directions, with fundamental contributions from Academicians L. I. Mandelshtam, N. D. Papalexi, V. A. Fok, A. I. Berg, V. A. Kotelnikov, and many other Soviet scientists.

As the reader knows from previous courses, the transmission of a message from the source to the recepient involves the use of a *communication channel*. Basically, a communication channel consists of a transmitter, a receiver, and the physical medium through which the emitted electromagnetic waves are propagated. This medium may be a free space or an engineering structure, such as waveguides, cables, and other transmission lines.

The *signal* coming from the originator at the sending end of a channel is converted by a microphone, a TV camera or other devices into electric oscillations. These oscillations cannot be used to excite electromagnetic waves in the medium directly because their frequency is too low. Therefore, the methods of signal

transmission used in communication are based on the fact that the low-frequency, or baseband, oscillations carrying the original message are utilized by suitable devices to vary the parameters of a sufficiently strong *carrier* whose frequency lies in the r.f. range. This form of signal transformation is called the *modulation* of the carrier.

The modulated signal is radiated by the antenna of the transmitter. The resultant electromagnetic waves produce in the antenna of a receiver a signal, usually at a very low power level. After frequency filtering and amplification, the received signal is subjected to *demodulation* (or *detection*) which is the reverse of modulation, in order to extract the original baseband signal from the modulated one. As a result, the signal appearing at the receiver's output is an exact replica of the transmitted original message.

From the above brief description of how a simple communication channel operates it is seen that the transmission of messages over the channel involves a variety of *signal transformations*. These transformations are implemented by appropriate physical systems known as (electric and electronic) *circuits*. Each circuit performs a particular transformation of the applied signal, and the nature of this transformation is wholly decided by the internal structure of the circuit. Accordingly, communication circuits are classed into amplifiers, frequency-selective filtering systems, waveform converters, modulators, detectors, and many other types considered in this course.

In any real communication channel, the wanted signal is always accompanied by interference arising from many factors, such as the noise created by the chaotic thermal motion of electrons in the circuit components, imperfect contacts in the apparatus, the effect of adjacent channels operating at closely spaced carrier frequencies, the presence of noise-producing cosmic rays, and the like. The ability of a communication system to transmit and receive information faithfully is known as *noise immunity*. The development of noise-immune systems is one of the principal tasks of present-day communication engineering. The theory and art of building such systems are the objectives of a separate division which has come to be known as *statistical communication* based on probabilistic methods. One of the most effective approaches to securing high noise immunity is the use of better forms of signal modulation, notably *optimal message coding*.

To sum up, the course on signals and circuits will be concerned with the following subjects:

1. The properties of various signals and interference and the principles underlying their mathematical description.

2. The properties of the physical systems acting as communication circuits.

3. The methods for the analysis of signal transformation in circuits, and for the synthesis of the basic types of circuits.
4. The synthesis of communication circuits having predetermined properties.

To-day, communication engineering is a rapidly growing and developing field of applied science. Speaking of its prospects in the nearest future, mention should be made of the tendency towards using electromagnetic waves at ever higher frequencies. For example, microwave frequencies utilized at one time solely in radar have now come to be widely used in television and telemetry. Impressive successes have been scored in the development of laser communication systems using carrier frequencies lying in the visible and infra-red regions of the electromagnetic spectrum.

Rapid progress has been registered in the field of circuit components and circuit design. Whereas traditional circuits are almost exclusively combinations of linear and nonlinear components, the more recent trend is towards the use of self-contained *functional devices and systems* which effect signal processing by virtue of the specific wave and oscillatory phenomena occurring in semiconductors, dielectrics and magnetics, that is, in the solid state. Microelectronic hardware, too, is playing an important role in present-day communication engineering. Readily available, reliable and lag-free, *integrated circuits* (ICs) have had an important impact on many fields of communication engineering. Microelectronics has been instrumental in the large-scale change-over to the fundamentally new *digital methods* of signal processing and transformation.

There is every reason to believe that all branches of communication will keep expanding and advancing on the basis of progress in many allied fields of science and technology.

1. Signals

Chapter 1
Elements of the General Theory of Signals

Along with matter and field, information belongs to the most important categories of the natural sciences

The term *signal* is frequently encountered not only in science and technology, but also in everyday life. Without giving a second thought to the matter, we sometimes do not differentiate the concepts of *signal*, *message*, and *information*. Ordinarily, this does not entail any confusion because the word signal has its origin in the Latin *signum* (sign) which has a broad range of senses. Still, in getting down to a systematic study of communication theory, we should refine the meaning of the concept of signal. By tradition, the term signal refers to time variations in the physical state of an object which serves to represent, register and transmit messages. In man's activities, messages are inseparably related to the information they carry.

The range of problems involving the concepts of message and information is very broad. They have long been drawing close attention from engineers, mathematicians, linguists, and philosophers. In the 1940s, C. Shannon laid the foundation for a new direction in science which has come to be known as information theory.

The problems that information theory has to do with usually lie far outside the scope of a course on signals and circuits. Therefore, this book will not dwell upon the relation between the physical form of the signal and the meaning of the message embedded in it. Nor will it discuss the value of the information embodied in a message and, in the final analysis, in a signal.

1.1 Classification of Signals

When one embarks on a study of some new objects or phenomena, one always strives to classify them in some preliminary manner. At this stage, our principal objective is to formulate criteria under which signals may be classified, and, which is very

important for the subsequent discourse, to define applicable terms.

A signal and its mathematical model. Signals as certain physical processes can be observed with a variety of devices and instruments, such as cathode-ray oscilloscopes, voltmeters, and receivers. This empirical approach suffers from a serious limitation. The phenomena examined by an experimenter always manifest themselves as specific, individual events which lack the degree of generality that could enable the experimenter to form an idea about their fundamental properties and to predict results under changed conditions.

For signals to serve as objects of a theoretical study and calculations, it is essential to tell how they can be described mathematically or, in scientific parlance, to build a *mathematical model* of the signal under study.

A mathematical model

A mathematical model of a signal is a functional relation in which the argument (or independent variable) is time. As a rule, we shall designate mathematical models of signals with letters of the English alphabet, as follows: $s(t)$, $u(t)$, $f(t)$, and so on.

The choice of a model is the first important step towards a systematic study of a phenomenon (in our case, a physical signal). Above all, a mathematical model enables the investigator to divorce himself from the specific nature of the signal carrier. In communication engineering, one and the same mathematical model can equally well describe current, voltage, electric field strength, and other quantities.

In most communication systems the signals are carried by electromagnetic oscillations and waves

Another aspect of the method based on mathematical modelling is that the investigator can limit himself solely to those of the properties of signals that are objectively most important and neglect a large number of secondary attributes of minor importance. For example, it would be extremely difficult in an overwhelming majority of cases to find exact functional relations which would fit the electric oscillations as they are observed experimentally. Yet, proceeding from what he knows about the system as a whole, the investigator can draw upon the multiplicity of mathematical models available to him and select one best suited to each particular case and possessing a maximum of simplicity. Thus, the choice of a model is in a sense a creative process.

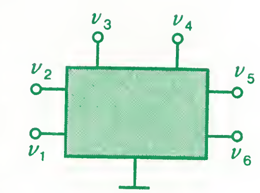
Real and complex signals

The functions describing signals can take on real and complex values. Quite aptly, in our discussion we shall frequently mention real and complex models of signals. Which to use is solely a matter of mathematical convenience.

Mathematical models of signals enable the investigator to compare them, to establish their similarity or difference, and, in the final analysis, to classify them.

One-dimensional and multidimensional signals. A typical communication signal is the voltage across the terminals of a circuit or the current flowing in that circuit. Such a signal can be described by one time function, so it is called a *one-dimensional* signal. In this text, we will be mostly concerned with one-dimensional signals. Sometimes, however, it will be convenient to consider *multidimensional* (or *vector*) signals of the form

$$\vec{V}(t) = [v_1(t), v_2(t), \ldots, v_N(t)]$$



formed by a set of one-dimensional signals. The number N is the *dimensionality* of a multidimensional signal. (The term dimensionality has been borrowed from linear algebra.)

An example of a multidimensional signal is the set of voltages existing at the terminals of a multiport.

Importantly, a multidimensional signal is an ordered set of one-dimensional signals. Therefore, signals differing in the sequence in which their components occur will not, in the general case, be identical:

$$\{v_1, v_2\} \neq \{v_2, v_1\}$$

The use of multidimensional signal models is especially warranted in cases where a complex system is being analyzed on a computer.

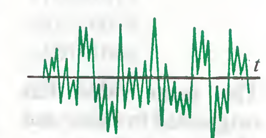
$v(t) = V_0 \cos \omega_0 t$ – a formula as a model of a deterministic signal

Deterministic and random signals. Alternatively, signals may be classified according as their instantaneous values can or cannot be predicted accurately at any instant of time. If the mathematical model of a signal permits this prediction, we have a *deterministic* signal. It can be specified in a variety of ways: by giving a mathematical formula, a computational algorithm, or simply its description in words.

▲ **Solve Problems 13 and 14**

Strictly speaking, deterministic signals are nonexistent in nature. The inevitable and unpredictable interaction of the message source with surrounding physical objects and the chaotic thermal fluctuations bring us to consider real signals as random or stochastic time functions, that is, to deal with *random* or *stochastic* signals.

In telecommunications, random oscillations frequently manifest themselves as noise interfering with the extraction of the valid information from the received oscillation. That is why the search for ways and means of controlling interference and improving the noise immunity of reception are pivotal problems in present-day communication engineering.



Waveform of a typical random signal

It may appear that the concept of random signal is contradictory. This is not so, however. For example, the signal appearing at the receiver output of a radio telescope trained on a source of cosmic radiation is composed of random fluctuations, but they carry variegated information about the natural object.

No hard and fast boundary exists between deterministic and random signals. Quite frequently, especially when the noise level is substantially lower than that of the wanted signal of known form, a simpler deterministic model may prove adequate.

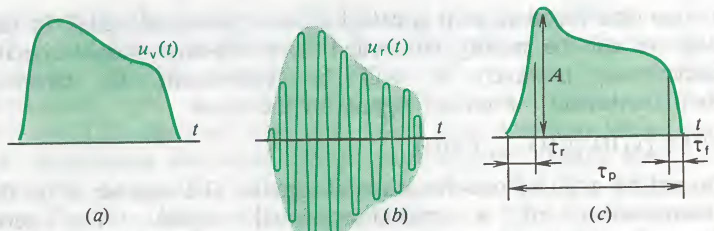


Fig. 1.1 Pulse signals and their characteristics: (*a*) video pulse; (*b*) radio pulse; (*c*) numerical parameters of a pulse

The techniques and procedures of statistical communication theory developed in recent decades to analyse the properties of random signals have quite a number of specific features and are based on the mathematical tools of probability theory and the theory of random processes. All this will take up the whole of Chapters 6 and 7.

A pulse

A video pulse and a radio pulse

The term video has originated in radar and television technology

Pulse signals. Pulses are a very important class of signals for telecommunications. By definition, a pulse is an oscillation which exists only within a finite span of time. It is customary to distinguish *video pulses* (Fig. 1.1*a*) and *radio pulses* (Fig. 1.1*b*). They differ in the following. If $u_v(t)$ is a video pulse, then the corresponding radio pulse will be

$$u_r(t) = u_v(t)\cos(\omega_0 t + \varphi_0)$$

where the frequency ω_0 and the initial phase φ_0 are arbitrary. In the above expression, $u_v(t)$ is the *envelope* of the radio pulse and the function $\cos(\omega_0 t + \varphi_0)$ is called the *carrier*.

Often, especially in engineering calculations, the complete mathematical model which accounts for all details in the "fine structure" of a pulse is replaced with the numerical parameters of the video pulse which give a simplified idea about its form. Thus, for a video pulse approaching a trapezoid in shape (Fig. 1.1*c*), it is convenient to know its *amplitude* (or height) A. Of the time parameters of a video pulse, its *duration*, τ_p, is the most important one. Frequently, it is also important to know the rise time, τ_r, and the fall (or decay) time, τ_f, of a pulse.

Communication engineering has to do with voltage pulses with amplitudes ranging from a few fractions of a microvolt to several kilovolts and with a duration of down to a few nanoseconds.

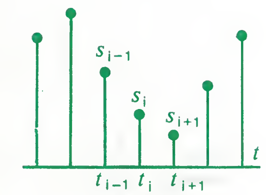
Continuous, discrete and digital signals. Before we conclude the brief overview of the principles underlying the classification of signals, the following should be noted.

Usually, any signal-producing physical process develops in time

in such a way that the signal value can be measured at any instant. Signals of this class are known as *continuous*. Another name is *analog* signals. The term analog is used since the waveform of the signal is analogous, or similar, to the input process waveform.

A one-dimensional continuous signal can be represented by its waveform (such as recorded on an oscillograph), which may be likewise continuous or discontinuous.

In the early days of communication services only continuous signals were used. Owing to their properties, they served well quite a number of tasks (radio communication, television, etc.). Also, continuous signals were simple to generate, receive and process by the facilities available at that time.



The model of a sampled (discrete) signal

As the requirements to be met by communication systems grew more stringent and the systems were called upon to perform an ever increasing number of tasks, the need arose to look for new design principles. In some cases, analog systems have given way to sampled-data (pulsed) systems which utilize signals produced by sampling. The simplest mathematical model of a *sampled (discrete) signal*, $s_d(t)$, is a countable set of points $\{t_i\}$ (where $i = 1, 2, 3, \ldots$) along the time axis, at each of which a sample, s_i, of the signal is determined. As a rule, the *sampling interval*, $\Delta = t_{i+1} - t_i$, is constant for a given signal.

An advantage which discrete (sampled) signals have over continuous signals is that there is no need for reproducing the signal continuously, at every instant along the time axis. Owing to this feature, one and the same communication system may be used to transmit messages from different sources to different users by means of what is known as time division multiplexing (TDM).

Obviously, in order to derive a sampled waveform from a continuous signal which rapidly varies in time a shorter sampling interval Δ is required. This fundamental matter will be taken up in more detail in Chapter 5.

. . . . . .
111001011
101110010
010011000
100110011
. . . . . .

Consecutive samples of a digital signal

Discrete signals include a special variety known as *digital signals*. They are called "digital" because their samples are represented by numbers (digits). For ease of engineering implementation and signal processing, binary numbers are used, with a limited and, as a rule, not very large number of digits or bits. Recently, there has been a growing trend towards using digital (sampled-data) systems on an ever wider scale. It has been stimulated by advances in microelectronics and integrated circuits.

It is important to remember that any discrete or digital signal (we mean a physical process rather than its mathematical model) is essentially a continuous signal. For example, a slowly varying continuous signal, $s(t)$, can be represented in a discrete (sampled) form as a sequence of rectangular video pulses of equal duration (Fig. 1.2*a*); the samples have amplitudes proportional to that of $s(t)$ at sampling points. Or we may proceed in a different way: we can

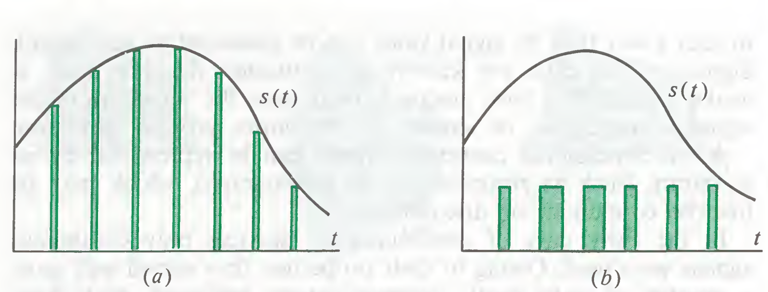
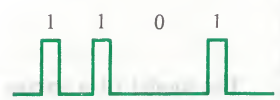


Fig. 1.2 Sampling of an analog signal: (*a*) with samples varying in height; (*b*) with samples varying in width (length)

hold the amplitude of the pulses at a constant value and vary their duration in proportion to the sample values (Fig. 1.2*b*).

What is important is that the two methods of converting a continuous signal into a sampled (discrete) signal are completely equivalent if we assume that the values of the continuous signal at the sampling points are proportional to the area of the individual video pulses.



In digital form, too, samples can be represented as a sequence of video pulses. The binary number system ideally suits the purpose. We may, for example, assign a high potential level to a 1, and a low potential level to a 0.

In more detail, discrete (sampled) signals and their properties will be studied in Chapter 15.

1.2 Dynamic Representation of Signals

Many tasks in telecommunications call for specific signal representations. Frequently it is essential to know not only the instantaneous (present) value of a signal, but also its behaviour along the time axis both in the past and in the future.

The principle of dynamic representation. Let us describe a real signal approximately, using a sum of elementary signals occurring at consecutive instants of time. By letting the duration of the individual elementary signals tend to zero, we shall, naturally, have an exact replica of the original signal in the limit. This is the *dynamic representation* of signals; the word "dynamic" emphasizes the fact that the process is varying in time.

Elementary signals may be chosen in any arbitrary manner, but two methods of dynamic representation are in a specially wide use. In one of them, elementary signals are step functions occurring at equal time intervals, Δ (Fig. 1.3*a*). The height of each step is equal to the change in the signal over the time interval Δ. In the other method, elementary signals are rectangular pulses. They are

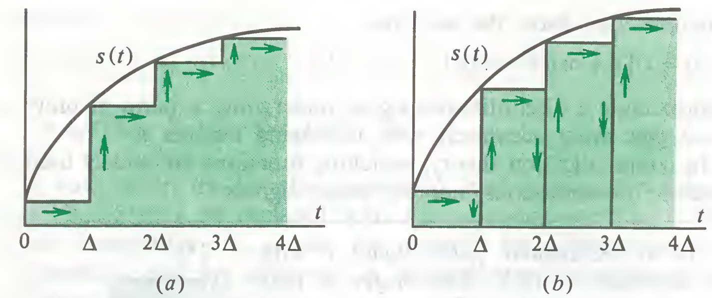
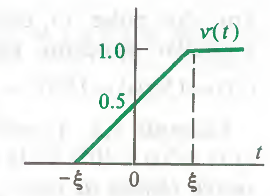


Fig. 1.3 Dynamic representation of signals. (The arrows indicate the direction of variation in the elementary components with time.)

contiguous to each other so that they form a sequence inscribed into or circumscribed around a curve (Fig. 1.3*b*).

Let us take a closer look at the properties of the elementary signal used in the first method of dynamic representation.



Switching function. Suppose the mathematical model of the signal is a set of equalities

$$v(t)=\begin{cases}0, & t<-\xi\\ 0.5(t/\xi+1), & -\xi\leqslant t<\xi\\ 1, & t>\xi\end{cases} \tag{1.1}$$

The function (1.1) describes the transition of a physical object from a "zero" to a "one" state, so that the transition is linear over the time interval 2ξ. If we let ξ tend to zero, then in the limit the transition will occur instantaneously. The mathematical model of this limiting signal has come to be known as the *switching function* or the *Heaviside function*:

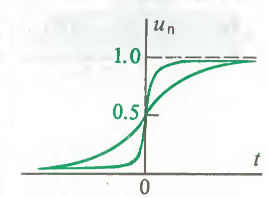
The switching or Heaviside function

$$\sigma(t)=\begin{cases}0, & t<0\\ 0.5, & t=0\\ 1, & t>0\end{cases} \tag{1.2}$$

Oliver Heaviside (1850-1925), a British physicist

Using the function $\sigma(t)$, it is convenient to describe various switching processes in electric circuits.

In the general case, the switching function may be translated from the origin of time by an amount t_0. The translated (or time-shifted) switching function is written as



$$\sigma(t-t_0)=\begin{cases}0, & t<t_0\\ 0.5, & t=t_0\\ 1, & t>t_0\end{cases} \tag{1.3}$$

The way the switching function is defined above is not the only one possible. As an example, which can readily be checked, the

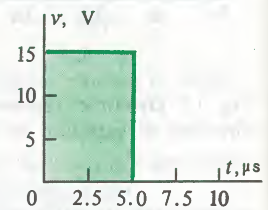
functions that form the sequence

$$u_n(t) = 1/[1 + \exp(-nt)]$$

approximate a discontinuous signal undergoing a jump of unity at $t = 0$ ever more accurately with increasing number n.

In communication theory, switching functions are widely used to describe discontinuous, notably pulse, signals.

Example 1.1. *Let there be a rectangular pulse signal v with a duration of 5 μs and an amplitude of 15 V. The origin of time coincides with the leading edge of the pulse. Write an analytical expression for this signal.*



The jump in level at $t = 0$ is described by the function

$$v = 15\sigma(t)$$

For the pulse to cease at $t = 5 \times 10^{-6}$ s, we should subtract a similar switching pulse delayed by this time interval so that

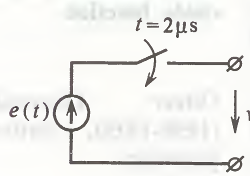
$$v(t) = 15\sigma(t) - 15\sigma(t - 5 \times 10^{-6})\ \text{V}$$

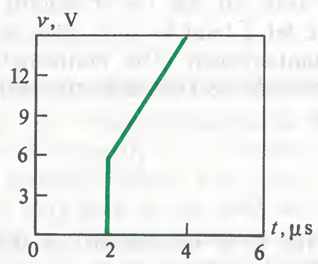
Example 1.2. *A source of an emf linearly varying with time as $e(t) = 3.0 \times 10^6 t$ V, is connected to an external circuit by a perfect switch closing at time $t_0 = 2$ μs. Write a mathematical model for the output voltage of the system.*

Work Problems 1 and 2

For times shorter than 2 μs, the output voltage of the source is zero. Therefore, it is obvious that

$$v(t) = 3.0 \times 10^6 t \times \sigma(t - 2 \times 10^{-6})\ \text{V}$$

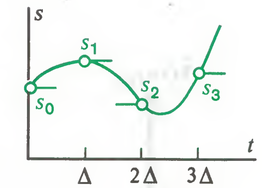




The above process may be written differently, if we represent it as the sum of the switching pulse occurring at the instant when the switch closes and a linearly rising pulse:

$$v(t) = [6 + 3 \times 10^6 (t - 2 \times 10^{-6})]\,\sigma(t - 2 \times 10^{-6})\ \text{V}$$

Dynamic representation of an arbitrary signal in terms of switching functions. Consider a signal $s(t)$, assuming that $s(t) = 0$ at $t < 0$. Let $\{\Delta, 2\Delta, 3\Delta, \ldots\}$ be a sequence of time instants and $\{s_1, s_2, s_3, \ldots\}$ be the corresponding sequence of signal values. If $s_0 = s(0)$ is the zeroth sample value, then, as follows from the plot, the value of the signal at any t is approximately equal to the sum of step functions



$$s(t) \approx s_0\sigma(t) + (s_1 - s_0)\sigma(t - \Delta) + (s_2 - s_1)\sigma(t - 2\Delta) + \ldots$$

$$= s_0\sigma(t) + \sum_{k=1}^{\infty} (s_k - s_{k-1})\sigma(t - k\Delta)$$

If, now, we let the sampling interval Δ tend to zero, the discrete variable $k\Delta$ may be replaced with a continuous variable τ. The small changes $(s_k - s_{k-1})$ may be replaced by differentials $ds = (ds/d\tau)d\tau$, and we arrive at a dynamic representation of an arbitrary signal in the form

$$s(t) = s_0\sigma(t) + \int_0^{\infty} (ds/d\tau)\,\sigma(t - \tau)\,d\tau \qquad (1.4)$$

Example 1.3. *The signal $s(t)$ is equal to zero at $t < 0$ and varies as a quadratic parabola, $s(t) = At^2$, at $t > 0$. Find the dynamic representation of this signal.*

Here, $s_0 = 0$, and $ds/d\tau = 2A\tau$. Therefore,

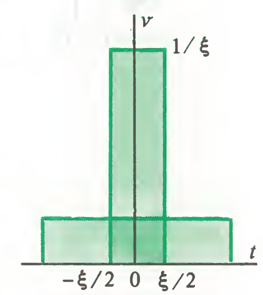
$$s(t) = 2A \int_0^{\infty} \tau\sigma(t - \tau)\,d\tau$$

The meaning of the last equation is that the height of the elementary steps that form the aggregate signal linearly rises with time.

In passing to the second method of dynamic representation of signals, when the expansion elements are short pulses, we should introduce a new important concept.

Delta-function. Consider a rectangular pulse signal defined by the formula

$$v(t; \xi) = (1/\xi)[\sigma(t + \xi/2) - \sigma(t - \xi/2)] \qquad (1.5)$$



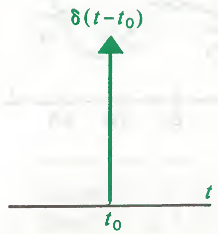
This pulse is characterized by the fact that with any choice of the parameter ξ it has unit area:

$$A_v = \int_{-\infty}^{\infty} v\,dt = 1$$

If, for example, v is a voltage, then $A_v = 1$ V s.

Now we let the parameter ξ tend to zero. While contracting in duration, the pulse retains its unit area. Therefore, its height must increase without bound. The limit of a sequence of such functions as $\xi \to 0$ is termed the *unit impulse function*, the *delta function*, or the *Dirac delta-function*:

$$\delta(t) = \lim_{\xi \to 0} v(t; \xi) \qquad (1.6)$$



The delta function in symbolic representation

The delta function is a very interesting mathematical entity. Being equal to zero everywhere except at the point $t=0$ (it is said to be concentrated at that point), this function does possess a unit integral

$$\int_{-\infty}^{\infty} \delta(t)\,dt = 1 \tag{1.7}$$

From a mathematical point of view, the Dirac delta-function is a *generalized function.* At present, the theory of generalized functions has gained a broad field of application. In many divisions of science, it has been used to study discontinuous processes that cannot be analysed by classical methods.

In this course, we shall widely use the concept of the Dirac delta-function. The main reason for the popularity of the delta function is this. As will be recalled from mechanics, if a variable force $F(t)$ acts on a material point of mass m over a time interval $[t_1, t_2]$, the change in the momentum of the point will be given by

$$mv_2 - mv_1 = \int_{t_1}^{t_2} F(t)\,dt$$

Thus, *what counts is not the force itself, but its momentum* appearing on the right-hand side of the above equation. The delta-function is just the mathematical model of a short input or stimulus of unit momentum (unit area).

Solve Problem 15

It is proved in mathematics that the properties of the delta function are inherent in many sequences of ordinary classical functions. We shall cite two typical examples [1]:

$$\delta(t) = \lim_{n\to\infty} \sqrt{n/2\pi}\,\exp(-nt^2/2) \tag{1.8}$$

$$\delta(t) = \lim_{n\to\infty} (\sin nt)/\pi t \tag{1.9}$$

Dynamic representation of signals in terms of delta functions. Let us go back to the description of a continuous signal as a sum of contiguous rectangular pulses (see Fig. 1.3*b*). If s_k is the kth sample of the signal, then the kth elementary pulse may be defined by the equation

$$\eta_k(t) = s_k[\sigma(t-t_k) - \sigma(t-t_k-\Delta)] \tag{1.10}$$

Under dynamic representation, the original signal $s(t)$ must be regarded as the sum of such elementary components:

$$s(t) = \sum_{k=-\infty}^{\infty} \eta_k(t) \tag{1.11}$$

At any t other than zero, this sum will contain only one kth term such that

$$t_k < t < t_{k+1}$$

On dividing and multiplying by the interval Δ, and substituting (1.10) into (1.11), we get

$$s(t) = \sum_{k=-\infty}^{\infty} s_k(1/\Delta)[\sigma(t-t_k) - \sigma(t-t_k-\Delta)]\Delta$$

In the limit $\Delta\to 0$, we must replace summation with integration with respect to the formal variable τ whose differential, $d\tau$, is analogous to Δ. Since

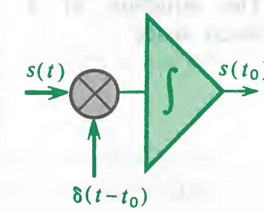
$$\lim_{\Delta\to 0} [\delta(t-\tau) - \sigma(t-\tau-\Delta)]/\Delta = \delta(t-\tau),$$

we obtain the sought-for formula for the dynamic representation of the signal

$$s(t) = \int_{-\infty}^{\infty} s(\tau)\,\delta(t-\tau)\,d\tau \tag{1.12}$$

An important property of the delta function is that *its physical dimensions are those of frequency.*

The filtering properties of the delta function



To sum up, if we multiply a continuous function by the delta function and integrate the product with respect to time, the result will be the value of the continuous function at the point where the δ-impulse is concentrated. In this lies, as is customary to say, the *filtering property of the delta-function.*

Among other things, hence arises the structure of a system for measuring the instantaneous values of a signal $s(t)$. The system must consist of two units, a multiplier and an integrator. Clearly, the accuracy with which the value $s(t_0)$ is determined will improve as the real signal (say, a rectangular video pulse) approximately representing the delta function becomes shorter.

1.3 Geometric Methods in Signal Theory

Many theoretical and applied problems of telecommunications involve questions like these: (1) In what sense may we say that one signal substantially exceeds another? (2) Is it possible to evaluate objectively how much two different signals "resemble" each other?

In the 20th century, mathematics has developed powerful techniques of functional analysis which generalize our intuitive ideas about the geometrical structure of space. As has been found, the concepts of functional analysis make it possible to formulate a consistent signal theory based on the treatment of a signal as a vector in a suitably constructed infinite-dimensional space.

Linear signal space. Let

$M = \{s_1(t), s_2(t), \ldots\}$

be a set of signals. The reason why these objects are combined is that each has properties common to all elements of the set M.

Example 1.4. *M is the set of all continuous signals which are nonzero in the time interval [0, 15 μs] and equal to zero outside that interval.*

Example 1.5. *M is a set of signals of the form*

$s_j(t) = A_j \cos(\omega_j t + \varphi_j)$

which are harmonic oscillations differing in amplitude, frequency and initial phase.

The properties of the signals that form a set can be investigated if it is possible to establish the relationship between the individual members of that set. Then, as is customary to say, the set possesses a definite *structure*. The choice of a particular structure for a set is dictated by physical considerations. For example, in the case of electric signals it is known that they can be combined and multiplied by an arbitrary scale factor. On this basis, we may use for such sets a structure known as a *linear space*.

The structure of a linear space

For a signal set M to form a real linear space, the following conditions (axioms) must be satisfied:

1. Any signal u which is an element of the set M, that is, $u \in M$, takes only real values for any t.

2. For any $u \in M$ and $v \in M$ there exists their sum

$w = u + v$

such that w also belongs to M, and the addition is both commutative

$u + v = v + u$

and associative

$u + (v + x) = (u + v) + x$

3. For any signal $s \in M$ and any real number α, there is a signal which is defined as

$f = \alpha s \in M$

4. The set M contains a unique member $\varnothing$ (called the origin) such that $u + \varnothing = u$ for each $u \in M$.

If we consider the mathematical models of signals taking on complex values and assume that in axiom (3) the multiplication is by a complex number, we shall arrive at the concept of a complex linear space.

The introduction of the linear space structure is the first step towards the geometric treatment of signals. The elements of linear spaces are frequently called vectors in order to stress their analogy in properties with conventional three-dimensional vectors.

Vectors

The constraints imposed by the axioms are very rigorous. In no way can just any set of signals be a linear space.

Example 1.6 *The set M is formed by rectangular voltage video pulses existing in the time interval [0, 20 μs], their amplitudes not exceeding 10 V. May this set be taken to be a linear space?*

If we add together pulses of amplitudes 6 V and 8 V, we will obtain a pulse which does not exist in the set M. Hence, M does not form a linear space.

The concept of a basis. As with a conventional three-dimensional space, a linear signal space may be given a special structure which plays the role of a coordinate system. It is then said that the set of vectors $\{e_1, e_2, e_3, \ldots\}$ belonging to M serves as a *linearly independent basis* for M, if the equality

Linear independence

$$\sum_i \alpha_i e_i = \varnothing \tag{1.13}$$

is satisfied only when the numerical coefficients α_i vanish all at one and the same time.

If we have an expansion of a signal $s(t)$ in the form

$s(t) = \sum_i c_i e_i$

then the numbers $\{c_1, c_2, \ldots\}$ are *projections* of the signal $s(t)$ relative to the adopted basis.

In signal theory, the number of basis vectors is, as a rule, infinitely large. Such linear spaces are called *infinite-dimensional.*

Example 1.7. *If a linear space is formed by signals which are described by polynomials of any order, however high:*

$$s(t) = \sum_{n=0}^{\infty} \beta_n t^n$$

(such functions are called analytic), then the basis for this space is a system of monomials

$\{e_0 = 1; e_1 = t; e_2 = t^2; \ldots\}$

A normed linear space. Signal energy. For further insight into the geometric treatment of signal theory, we need one more concept which would correspond in meaning to the length of a vector. With it, we cannot only give a more exact meaning to a statement such

as "The first signal is larger than the second one", but also state by how much it is larger.

The norm of a signal

In mathematics, the analogue of a vector's length is its *norm*. A linear signal space L is said to be *normed*, if to each vector $s(t)\in L$ is assigned a unique real number $\|\vec{s}\|$. This number is called the norm of a vector; it satisfies the following axioms of a normed space:

1. The norm is non-negative, that is, $\|\vec{s}\| \geqslant 0$, with $\|\vec{s}\| = 0$ if and only if $s = \varnothing$.

2. For any number α the equality

$$\|\alpha\vec{s}\| = |\alpha| \cdot \|\vec{s}\|$$

is satisfied.

3. If $s(t)$ and $p(t)$ are two vectors out of the set L, the triangle inequality

$$\|\vec{s}+\vec{p}\| \leqslant \|\vec{s}\| + \|\vec{p}\|$$

is satisfied.

The norm of a signal may be defined in more than one way. In communication theory, it is most often assumed that real signals have the norm

$$\|\vec{s}\| = \sqrt{\int_{-\infty}^{\infty} s^2(t)\,dt} \tag{1.14}$$

(of two likely values of the root, we choose the positive one). For complex signals

The energy of a signal

$$\|\vec{s}\| = \sqrt{\int_{-\infty}^{\infty} ss^*\,dt} \tag{1.15}$$

where s^* is the complex conjugate of s.

The square of the norm is termed the *signal energy*:

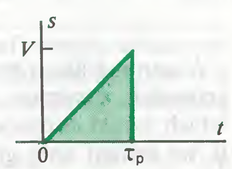
If a signal represents a voltage, the units of signal energy are V^2 s

$$E_s = \|\vec{s}\|^2 = \int_{-\infty}^{\infty} s^2(t)\,dt \tag{1.16}$$

Exactly this amount of energy will be dissipated in a 1-Ω resistor if $s(t)$ is the voltage applied across its terminals.

Example 1.8. *Find the energy and norm of the signal $s(t)$ which is a triangular voltage pulse of height V and of duration τ_p.*

In the time interval $[0, \tau_p]$, the signal is described by the function



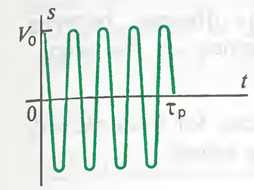
$$s(t) = Vt/\tau_p$$

The energy of the signal is

$$E_s = (V^2/\tau_p^2)\int_0^{\tau_p} t^2\,dt = V^2\tau_p/3$$

The norm of the signal is

$$\|\vec{s}\| = \sqrt{E_s} = V\sqrt{\tau_p}/\sqrt{3}$$



Example 1.9. *Find the energy of a radio pulse with a rectangular envelope. The pulse exists on the time interval* $[0, \tau_p]$ *and is described by the function*

$$s(t) = V_0 \cos(\omega_0 t + \varphi_0)$$

According to Eq. (1.16),

$$E_s = V_0^2 \int_0^{\tau_p} \cos^2(\omega_0 t + \varphi_0)\,dt = (V_0^2/\omega_0) \int_0^{\omega_0\tau_p+\varphi_0} \cos^2\eta\,d\eta$$

$$= (V_0^2/4\omega_0)[2(\omega_0\tau_p + \varphi_0) + \sin 2(\omega_0\tau_p + \varphi_0)]$$

If the pulse duration is many times the period of the carrier frequency, that is, if $\omega_0\tau_p \gg 1$, then

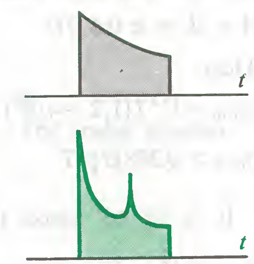
$$E_s \approx V_0^2\tau_p/2$$

irrespective of the values of ω_0 and φ_0.

It is convenient to determine the norm of a signal by Eq. (1.15) for the following reasons.

1. In communication engineering it is customary to state the magnitude of a signal in terms of the overall power effect, for example, the quantity of heat dissipated in a resistor.

2. The energy norm thus introduced is "insensitive" to variations in the signal waveform, which may be substantial, but occurring over relatively short spans of time.



These signals only slightly differ in energy

The normed linear space in which all vectors have a finite norm is termed the square-integrable function space and is symbolized as L_2.

Metric space. Let us introduce one more fundamental concept which will generalize our idea about the distance between points in space.

The metric

It is customary to say that a linear space L becomes metric, if to each pair of elements belonging to that space, $u, v \in L$, is assigned a real, positive number $\rho(u, v)$, called the *metric*, or the *distance* between those elements. Whatever the method used to define it, the metric must obey the axioms of a metric space:

1. $\rho(u, v) = \rho(v, u)$ (reflexivity of the metric).
2. $\rho(u, v) = 0$ for any $u \in L$.
3. Whatever the element $w \in L$, it is always that

$$\rho(u, v) \leqslant \rho(u, w) + \rho(w, v)$$

Ordinarily, the metric is defined as the norm of the difference between two signals:

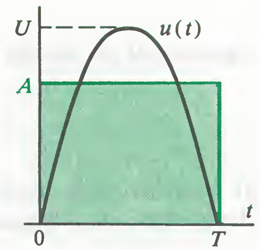
$$\rho(u, v) = \| u - v \| \tag{1.17}$$

In turn, the norm may be construed as the distance between a selected element of the space and the null element or the origin:

$$\| u \| = \rho(u, \varnothing)$$

By drawing upon the concept of metric, we can, for example, say how well one of the signals approximates the other.

Example 1.10. *The signal $u(t)$ is a chopped sinusoid falling to zero at the ends of the interval $[0, T]$. The pulse height U is known. Find the amplitude A of a rectangular pulse $v(t)$ of the same duration, such that the distance between the two signals is a minimum.*



The signal $u(t)$ is defined by the formula

$$u(t) = U \sin(\pi t/T), \quad 0 < t < T$$

The square of the distance between the signals is

$$\rho^2(u, v) = \int_0^T [U \sin(\pi t/T) - A]^2 \, dt = U^2 T/2 - 4AUT/\pi + A^2 T$$

A test of the above expression for an extremum shows that the distance will be a minimum if

$$A = 2U/\pi \approx 0.637U$$

Then,

$$\rho^2_{\min} = U^2 T(1/2 - 4/\pi^2) \approx 0.095 U^2 T$$

$$\rho_{\min} \approx 0.308\, U \sqrt{T}$$

▲ Work Problem 9

It is to be noted that the energy of the sinewave pulse is

$$E_v = U^2 \int_0^T \sin^2(\pi t/T)\, dt = U^2 T/2$$

and its norm is

$$\| u \| \approx 0.707 U \sqrt{T}$$

Thus, in terms of the metric we have chosen, the minimal attainable distance between the signals in question is 44% of the norm of the sinewave signal.

1.4 The Theory of Orthogonal Signals

Although we have introduced the linear space structure in a signal set, and defined the norm and the metric, we are still not able to find an important characteristic – the angle between two

vectors. This can be done, if we introduce the concept of the scalar product between the elements of a linear space.

Scalar product of signals. As will be recalled, if in a conventional three-dimensional space we know two vectors, $\vec{A}$ and $\vec{B}$, the square of the modulus of their sum is

$$|\vec{A} + \vec{B}|^2 = |\vec{A}|^2 + |\vec{B}|^2 + 2(\vec{A}\vec{B}) \tag{1.18}$$

where

$$(\vec{A}\vec{B}) = |\vec{A}| \cdot |\vec{B}| \cos \psi$$

is the scalar product of these vectors, dependent on the angle ψ between them.

By analogy, we can find the energy of the sum of two signals, u and v:

$$E_\Sigma = \int_{-\infty}^{\infty} (u + v)^2 \, dt = E_u + E_v + 2 \int_{-\infty}^{\infty} uv \, dt \tag{1.19}$$

In contrast to the signals themselves, their energies are *nonadditive – the energy of the aggregate signal contains what is called the cross energy*

● The cross energy

$$E_{uv} = 2 \int_{-\infty}^{\infty} uv \, dt$$

From comparison of Eqs. (1.18) and (1.19), we can define the *scalar product of the signals* u and v as

● The scalar product

$$(u, v) = \int_{-\infty}^{\infty} u(t) v(t) \, dt \tag{1.20}$$

The cosine of the angle between them is

$$\cos \psi_{uv} = \frac{(u, v)}{\| u \| \cdot \| v \|} \tag{1.21}$$

A scalar product possesses the following obvious properties:

(1) $(u, u) \geqslant 0$;

(2) $(u, v) = (v, u)$;

(3) $(\lambda u, v) = \lambda (u, v)$, where λ is a number; (1.22)

(4) $(u + v, w) = (u, w) + (v, w)$.

● The Hilbert space

The linear space containing the scalar product defined in Eq. (1.20) and satisfying the conditions (1.22) is called the *real Hilbert space*, H.

It is proved in mathematics that the Hilbert space satisfies the

David Hilbert (1862-1943), a prominent German mathematician

fundamental Cauchy-Buniakovski inequality

$$|(u, v)| \leqslant \|u\| \cdot \|v\| \tag{1.23}$$

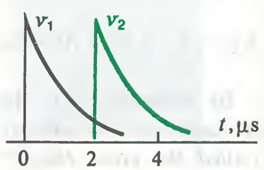
If the signals take on complex values, we may define a *complex Hilbert space* by introducing the scalar product according to the equation

$$(u, v) = \int_{-\infty}^{\infty} u(t)v^*(t)\,dt \tag{1.24}$$

Example 1.11. *There are two exponential voltage pulses (measured in volts, V) translated in time relative to each other and defined as follows:*

$$v_1(t) = 5\exp(-10^5 t)\sigma(t)$$

$$v_2(t) = 5\exp[-10^5(t - 2\times 10^{-6})]\sigma(t - 2\times 10^{-6})$$



Find the scalar product of the signals and the angle between them.

The two signals have the same energy:

$$\|v_1\|^2 = \|v_2\|^2 = 25\int_0^{\infty} \exp(-2\times 10^5 t)\,dt$$

$$= 1.25\times 10^{-4}\ \text{V}^2\,\text{s}$$

Their scalar product is

$$(v_1, v_2) = 25\int_0^{\infty} \exp(-10^5 t)\exp[-10^5(t + 2\times 10^{-6})]\,dt$$

$$= 1.023\times 10^{-4}\ \text{V}^2\,\text{s}$$

Hence,

$$\cos\psi_{v_1 v_2} = 0.819$$

and

$$\psi_{v_1 v_2} = 35°$$

Orthogonal signals and generalized Fourier series. Two signals, u and v, are called *orthogonal* if their scalar product and, in consequence, their cross energy are zero:

The principle of orthogonality

$$(u, v) = \int_{-\infty}^{\infty} u(t)v(t)\,dt = 0 \tag{1.25}$$

Figuratively speaking, orthogonal signals "do not look alike" as any pair of signals could.

Let H be a finite-energy signal Hilbert space. The signals are defined over a time interval $[t_1, t_2]$, which may be finite or infinite. Suppose that over the same time interval we define an infinite system of functions $\{u_1, u_2, \ldots, u_n, \ldots\}$ which are pairwise mutually orthogonal and have unit norm:

$$(u_i u_j) = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \tag{1.26}$$

It is then said that an *orthonormal basis* is specified in the signal space.

An orthonormal basis

Let us expand an arbitrary signal $s(t) \in H$ into a series:

$$s(t) = \sum_{i=1}^{\infty} c_i u_i(t) \tag{1.27}$$

The representation in (1.27) is called a *generalized Fourier series* of the signal $s(t)$ in the selected basis.

The coefficients of this series are found as follows. We multiply both sides of the equality (1.27) by an arbitrary kth basis function, u_k and integrate the results with respect to the time interval over which the signals are defined:

$$\int_{t_1}^{t_2} s(t)u_k(t)\,dt = \sum_{i=1}^{\infty} c_i \int_{t_1}^{t_2} u_i u_k\,dt \tag{1.28}$$

Since the basis has the property of orthonormality, the right-hand side of (1.28) retains only one, kth, element of the sum. Therefore,

An algorithm for finding the coefficients of a generalized Fourier series

$$c_k = \int_{t_1}^{t_2} s(t)u_k(t)\,dt = (s, u_k) \tag{1.29}$$

That a signal can be represented in terms of a generalized Fourier series is a fundamentally important factor. Rather than study a functional relationship at an uncountable set of points, we can characterize this signal with a countable (although, generally speaking, infinite) system of coefficients of a generalized Fourier series, c_k, which are projections of the vector $s(t)$ in a Hilbert space onto the basis directions.

Examples of orthonormal bases. How infinite systems of mutually orthogonal functions can be constructed is examined in detail in mathematics (see, for example, [2]). Here, we shall only give as examples two of the most important and commonly encountered cases.

An orthogonal set of harmonic signals. The reader can prove himself that over the interval $[0, T]$ a set of trigonometric functions of multiple frequencies, extended to include a time-invariant signal:

$$
\begin{aligned}
u_0 &= 1/\sqrt{T}\\
u_1 &= \sqrt{2/T}\sin 2\pi t/T\\
u_2 &= \sqrt{2/T}\cos 2\pi t/T \qquad (1.30)\\
&\cdots\cdots\cdots\\
u_{2m-1} &= \sqrt{2/T}\sin 2\pi m t/T\\
u_{2m} &= \sqrt{2/T}\cos 2\pi m t/T\\
&\cdots\cdots\cdots
\end{aligned}
$$

constitutes an orthonormal basis.

The expansion of periodic functions into a series in terms of the above basis will be taken up in the next chapter.

Signals corresponding to Walsh functions can be readily generated by microelectronic switches

Walsh functions. In recent time, advances in the processing of discrete signals have given impetus to the use of orthonormal Walsh functions. A distinction of these functions is that over the interval of their existence, $[-T/2, T/2]$, they take on only two values ($+1$ and -1) which, as is seen, differ solely in sign.

Let us introduce a dimensionless time $\vartheta = t/T$ and designate the kth Walsh function by the symbol wal (k, ϑ). These functions are far more difficult to specify analytically (see Appendix 1). The basic idea, however, can readily be gleaned from Fig. 1.4 which shows plots of the first few Walsh functions.

Interestingly, the kth Walsh function has the number k of zero crossings (sign changes) over the existence interval.

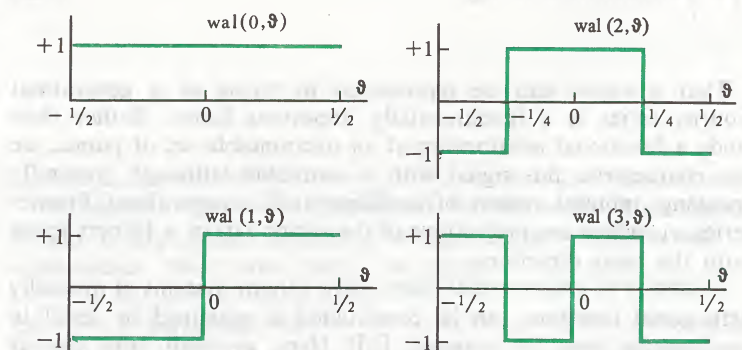


Fig. 1.4 Plots of the first few Walsh functions

Obviously, the condition of normality for Walsh functions for any value of k is

▲ Solve Problem 11

$$\|\mathrm{wal}(k, \vartheta)\|^2 = \int_{-1/2}^{1/2} \mathrm{wal}^2(k, \vartheta)\,d\vartheta = 1$$

The orthogonality of these functions is assured by the principle of construction and can be verified directly. For example,

$$\int_{-1/2}^{1/2} \mathrm{wal}(1, \vartheta)\,\mathrm{wal}(2, \vartheta)\,d\vartheta = \int_{-1/2}^{-1/4} (-1)^2\,d\vartheta + \int_{-1/4}^{0} (-1)\times 1\,d\vartheta + \int_{0}^{1/4} 1^2\,d\vartheta + \int_{1/4}^{1/2} 1\times(-1)\,d\vartheta = 0$$

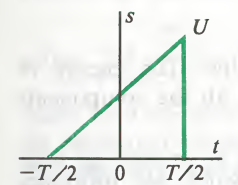
The expansion of a finite-energy signal defined over the time interval $[-T/2, T/2]$ into a generalized Fourier series in terms of Walsh functions, has the form

$$s(\vartheta) = \sum_{k=0}^{\infty} c_k \mathrm{wal}(k, \vartheta);\ \ \vartheta = t/T \qquad (1.31)$$

Example 1.12. *Find the first two coefficients in the expansion of a triangular pulse in terms of Walsh functions.*

Over the time interval $[-T/2, T/2]$, the signal is described by a function of the form

$$s(t) = U(t/T + 1/2)$$

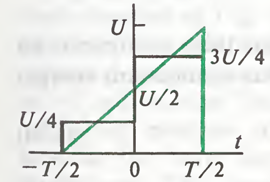


The coefficients of the generalized Fourier series are:

$$c_0 = \int_{-1/2}^{1/2} s(\vartheta)\,\mathrm{wal}(0, \vartheta)\,d\vartheta = U\int_{-1/2}^{1/2} (\vartheta + 1/2)\,d\vartheta = U/2$$

$$c_1 = \int_{-1/2}^{1/2} s(\vartheta)\,\mathrm{wal}(1, \vartheta)\,d\vartheta$$

$$= -U\int_{-1/2}^{0} (\vartheta + 1/2)\,d\vartheta + U\int_{0}^{1/2} (\vartheta + 1/2)\,d\vartheta = U/4$$



Thus, when a triangular pulse is approximated by the first two members of a Walsh function series, the result is a step function. Interestingly, even this rough approximation is satisfactory from the view-point of the energy norm introduced earlier. To demonstrate,

$$E_s = U^2\int_{-1/2}^{1/2} (\vartheta + 1/2)^2\,d\vartheta = U^2/3$$

whereas the difference energy

$$\| s(\vartheta) - c_0 \,\mathrm{wal}(0, \vartheta) - c_1 \,\mathrm{wal}(1, \vartheta) \|^2 = 4U^2 \int_0^{1/4} \xi^2 \, d\xi = \frac{U^2}{3 \times 16}$$

is only one-sixteenth, or 6.25%, of the energy of the signal being approximated.

The energy of a signal represented by a generalized Fourier series. Consider a signal $s(t)$ expanded into a series in terms of orthonormal basis functions:

$$s(t) = \sum_{k=1}^{\infty} c_k u_k(t)$$

and determine its energy:

$$E_s = \int_{t_1}^{t_2} s^2 dt = \int_{t_1}^{t_2} \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} (c_i c_j) u_i u_j dt$$

$$= \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} (c_i c_j) \int_{t_1}^{t_2} u_i u_j dt \tag{1.32}$$

Because the basis functions are orthonormal, only those terms in the sum of (1.32) are non-zero for which $i = j$. To sum up, we have obtained a remarkable result:

This equation generalizes Pythagoras' theorem to an infinite-dimensional space

$$E_s = \sum_{i=1}^{\infty} c_i^2 \tag{1.33}$$

The meaning of the above expression is this: The energy of a signal is equal to the sum of the energies of all the components that make up the generalized Fourier series.

Optimality of the signal expansion in terms of orthogonal basis functions. Let there be a signal, $s(t)$, for which we introduce a finite-dimensional approximation

$$\tilde{s}(t) = \sum_{k=1}^{N} c_k u_k(t)$$

with unknown coefficients c_k, and require that these coefficients be chosen such that the approximation error has a minimum energy:

$$\mu = \| s - \tilde{s} \|^2 = \int_{t_1}^{t_2} [s - \sum_k c_k u_k]^2 dt = \min \tag{1.34}$$

The necessary minimum condition consists in that the coefficients

c_m must satisfy a set of linear equations

$$\partial \mu / \partial c_m = 0, \quad m = 1, 2, \ldots, N \tag{1.35}$$

In the expanded form, the approximation error is

$$\mu = \int_{t_1}^{t_2} [s^2 - 2s \sum_{k=1}^{N} c_k u_k + \sum_{i=1}^{N} \sum_{j=1}^{N} c_i c_j u_i u_j] dt$$

Since the basis functions involved are orthogonal, it follows that

$$\frac{\partial}{\partial c_m} \left(\int_{t_1}^{t_2} [c_m^2 u_m^2 - 2 s c_m u_m] dt \right) = 0$$

In view of the fact that the basis functions have a norm of unity, we may conclude that equalities (1.35) will be satisfied when the expansion coefficients are chosen such that

$$c_m = \int_{t_1}^{t_2} s(t) u_m(t) dt$$

which fully checks with Eq. (1.29) for the coefficients of a generalized Fourier series.

A more thorough analysis (when one considers not only the first, but also the second derivative of the error energy) will show that the expansion of a signal into a generalized Fourier series assures a minimum of the approximation error.

It is to be noted in conclusion that, *by definition*, the Hilbert signal space possesses the important property of *completeness* which consists in the following. If the limit of the sum

The completeness of space

$$s(t) = \lim_{N \to \infty} \sum_{i=1}^{N} c_i u_i(t)$$

exists, then the limit itself is always an element of the Hilbert space. An important property of this class of spaces is that the norm of the approximation error is a monotonically decreasing function of N, the number of expansion terms considered. By choosing N sufficiently large, it is always possible to minimize the error to any acceptably small value.

Implementation of orthogonal signal expansion. Turn to the block diagram of Fig. 1.5 which shows a set-up used to determine experimentally the expansion coefficients for an arbitrary signal expanded into a generalized Fourier series in terms of a specified set of orthonormal basis functions.

The key elements of the set-up are function generators which generate the basis functions in terms of which the expansion is carried out. The signal in question is simultaneously fed to an assemblage of multipliers each of which multiplies the applied

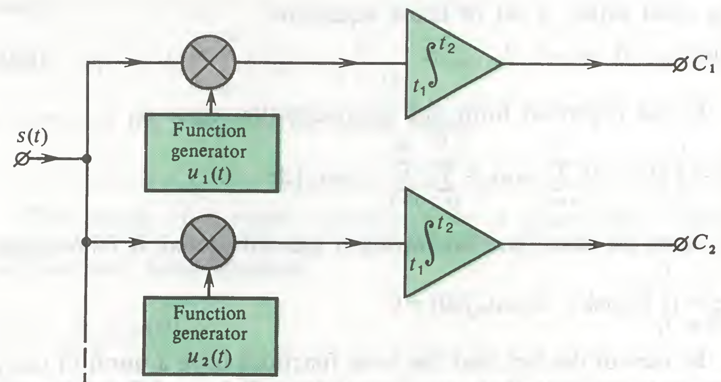


Fig. 1.5 Block diagram of a signal analyzer

signal by an appropriate basis function. From the multiplier outputs, the resultant signals are applied to integrators.

With this form of signal processing, a time-invariant signal appears across the output of each multiplier at the end of the integration time. The magnitude of the output signal is, in accord with Eq. (1.29), exactly equal to a particular coefficient of the generalized Fourier series.

Clearly, the performance of the set-up as a whole depends on how accurately the basis functions can be reproduced and also on how well the multipliers and the integrators operate.

The set-up shown in Fig. 1.5 is important not only in an applied but also in a theoretical sense. From its analysis we can see that all of the information embedded in the signal can be represented as a set of numbers, although infinite, but yet countable.

Summary

- For a theoretical study of signals, it is necessary to build their mathematical models.
- Signals are classified on the basis of the significant attributes of the corresponding mathematical models. It is customary to class signals into one-dimensional and multi-dimensional, deterministic and random, continuous and discrete. There is a special variety of discrete signals, called digital signals.
- In dynamic representation, signals can be described, taking into account their behaviour both in the "past" and in the "future".
- Dynamic representation utilizes two elementary signals, one being the switching function and the other, the Dirac delta-function.
- By introducing appropriate structures, some sets of signals can be transformed into linear functional spaces.

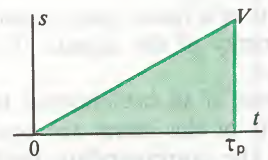
- A system of linearly independent vectors forms a basis in terms of which an arbitrary element belonging to a linear space can be expanded.
- In a linear signal space, the analogue of the length of a vector is its norm. The square of the norm is called the energy of the signal.
- A linear signal space turns into a metric space if one can determine the distance (metric) between two vectors.
- In order to determine the angle between two elements of a linear space, resort is made to their scalar product proportional to the cross energy of the signals. If the scalar product is zero, the signals are called orthogonal.
- The representation of a signal as an expansion in terms of an orthonormal basis set is termed a generalized Fourier series. The coefficients of this series are equal to the scalar products of the signal being expanded and the corresponding basis vectors.
- The most important examples of orthonormal basis sets are harmonic functions at multiple frequencies and Walsh functions.
- The energy of a signal is equal to the sum of all the components that form the generalized Fourier series.
- The expansion of a signal in terms of orthonormal basis functions yields a minimal r.m.s. error of approximation.
- The extraction of useful information from a signal may be visualized as an experimental evaluation of the generalized Fourier series coefficients of that signal.

Review Questions

1. Name two or three examples of physical processes describable with stochastic mathematical models.
2. Name the characteristics used to describe pulse signals.
3. Define the difference between a video pulse and a radio pulse.
4. Under what conditions will the replacement of a continuous signal with a discrete signal be inadequate?
5. Formulate the principle of dynamic signal representation.
6. List the main properties of the delta function.
7. State the most important axioms of a linear space.
8. What is the physical meaning of the square of the signal norm?
9. What is the geometric meaning of the Cauchy-Buniakovski inequality?
10. Draw several plots illustrating orthogonal signals.
11. What is the Hilbert space?
12. What makes the expansion of signals in terms of orthogonal Walsh functions convenient?

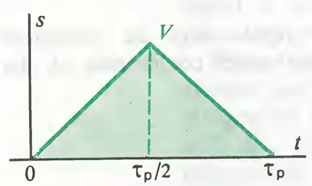
Problems

1. The figure shows a triangular voltage pulse.

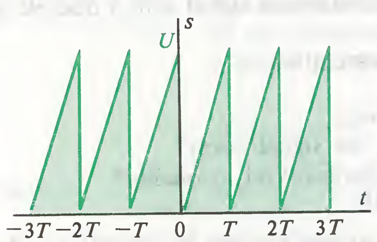


Write a mathematical model for this signal, using a combination of switching functions.

2. Solve Problem 1 for a symmetrical triangular pulse.



3. Write a mathematical model for an infinite sequence of identical triangular pulses.



4. Using Eq. (1.4), find a dynamic representation for the exponential video pulse described by the formula:

$u(t) = U \exp(-\alpha t)\,\sigma(t)$

5. Using Eq. (1.4), find a dynamic representation for the Gaussian video pulse

$u(t) = U \exp(-\beta t^2)$

defined along the entire infinite time axis. Note the way in which Eq. (1.4) must be modified.

6. Show that the delta-function can be construed as the derivative of the switching function

$\delta(t) = d\sigma/dt$

Hint. Take the derivative of the function $v(\xi, t)$ represented by Eq. (1.1), and pass to the limit with $\xi \to 0$.

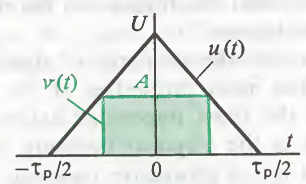
7. Find the energy and norm of the signal described by a mathematical model of the form

$u(t) = U_0 \exp(-\alpha |t|)$

8. Find the energy and norm of the cosinusoidal pulse

$$s(t) = \begin{cases} 0, & \omega_0 t < -\pi/2 \\ U\cos\omega_0 t, & -\pi/2 < \omega_0 t < \pi/2 \\ 0, & \omega_0 t > \pi/2 \end{cases}$$

9. The signal $u(t)$ is a symmetrical triangular pulse; the signal $v(t)$ is a rectangular pulse inscribed in the first.



Find the amplitude of the rectangular pulse such that the distance between the two pulses is a minimum.

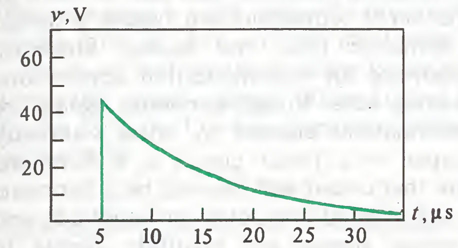
10. Using the principle of orthogonality, construct a plot of the function wal (4, ϑ) on the basis of Fig. 1.4.

11. Show that the distances between any two functions out of the set wal (0, ϑ), wal (1, ϑ) and wal (2, ϑ) are the same and equal to $\sqrt{2}$.

12. Apply a similar analysis to an orthonormal set of trigonometric functions (see Eq. (1.30)). Compare the results. Is it possible to draw an analogy with Pythagoras' theorem?

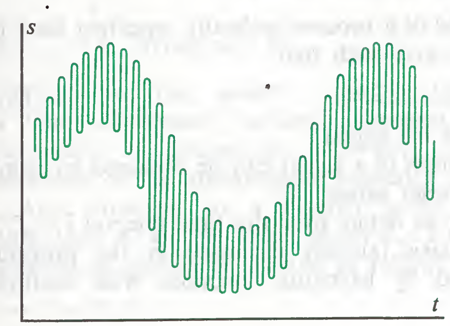
Advanced Problems

13. An experiment has yielded the following oscillogram of a signal:

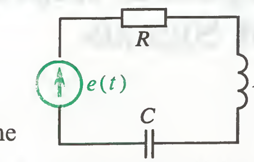


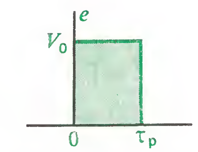
It is hypothesized that the signal can be described by an exponential time function. Propose the simplest possible graphical method to test this hypothesis.

14. Propose a mathematical model for the following signal:



15. A series resonant circuit is driven by a source of pulsed emf. The system's parameters are: $R = 5\ \Omega$, $L = 10\ \mu H$, and $C = 2$ nF. Pulse duration is $\tau_p = 0.5\ \mu s$, and pulse amplitude is $V_0 = 12$ V.





Prove that the real pulse can in this case be replaced with a mathematical model of the form $A\delta(t)$. What should the coefficient A be in this case?

List several simple physical situations out of everyday practice, when the actual effect on a system may be approximated with a delta-impulse.

16. Prove that the delta-function may be treated as the limit

$$\delta(t) = \lim_{q\to\infty}\left(\frac{1}{\pi}\,\frac{\sin qt}{t}\right)$$

17. Generalize the concepts of energy and norm to vector signals of an arbitrary dimensionality N.

18. Prove that if H is a real Hilbert space containing the signals u and v, the parallelogram equality is valid:

$$\|u+v\|^2 + \|u-v\|^2 = 2\|u\|^2 + 2\|v\|^2$$

19. Prove that the identity

$$4(u, v) = \|u+v\|^2 - \|u-v\|^2 + j\|u + jv\|^2 - j\|u - jv\|^2$$

holds in a complex Hilbert space.

20. Let $\{u_k(t),\ k = 1, 2, \ldots\}$ be an orthonormal basis in a Hilbert space H. Prove that Parseval's theorem

$$(s_1, s_2) = \sum_{k=1}^{\infty} (s_1, u_k)(s_2, u_k)$$

holds for any $s_1, s_2 \in H$.

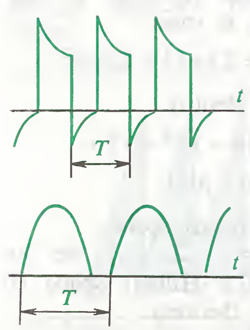
Chapter 2

Spectral Representations of Signals

Of the many sets of orthogonal functions that can be used as basic ones in the representation of communication signals, a special place is occupied by harmonic (sine and cosine) functions. Harmonic signals are important for communication applications due to several reasons. Among other things, harmonic signals are invariant under the transformations effected by linear stationary electric circuits. If the input to a linear circuit is a harmonic oscillation, the output from that circuit will likewise be a harmonic oscillation differing from the original one solely in amplitude and initial phase. Also, harmonic signals are relatively simple to generate.

If a signal is represented as the sum of harmonic oscillations differing in frequency, the signal is said to have been resolved into its spectrum in terms of harmonic functions. The sum of the individual harmonic components of the signal forms its *spectrum*.

2.1 Periodic Signals and Fourier Series



Examples of periodic signals

The mathematical model of a process cyclically repeating itself in time is a periodic signal, $s(t)$, such that

$$s(t) = s(t + nT) \qquad n = \pm 1, \pm 2, \ldots \tag{2.1}$$

where T is the period of the signal.

The spectral representation of a signal can be obtained by using the expansion into a Fourier series.

The Fourier series. Let us define over the time interval $[-T/2, T/2]$ the orthonormal basis (already discussed in the previous chapter) which is formed by harmonic functions with multiple frequencies:

$$\begin{aligned} u_0 &= 1/\sqrt{T} \\ u_1 &= \sqrt{2/T}\sin 2\pi t/T \\ u_2 &= \sqrt{2/T}\cos 2\pi t/T \\ u_3 &= \sqrt{2/T}\sin 4\pi t/T \\ u_4 &= \sqrt{2/T}\cos \pi t/T \\ &\ldots\ldots\ldots\ldots \end{aligned} \tag{2.2}$$

Any function u_m out of this basis satisfies the condition of periodicity, Eq. (2.1). On expanding the signal $s(t)$ in terms of this basis, that is, on finding the coefficients

$$c_m = (s, u_m) \tag{2.3}$$

we get its spectral representation

$$s(t) = \sum_{m=0}^{\infty} c_m u_m(t) \tag{2.4}$$

The expansion is valid along the entire infinite time axis. The series of the form in (2.4) is called the *Fourier series*.

The fundamental frequency

Let $\omega_1 = 2\pi/T$ be the *fundamental frequency* of the sequence that forms the periodic signal. When finding the expansion coefficients by Eq. (2.3), the Fourier series for a periodic signal may be written in the form

$$s(t) = a_0/2 + \sum_{n=1}^{\infty} (a_n \cos n\omega_1 t + b_n \sin n\omega_1 t) \tag{2.5}$$

where the coefficients are

$$\begin{aligned} a_0 &= \frac{2}{T}\int_{-T/2}^{T/2} s(t)\,dt \\ a_n &= \frac{2}{T}\int_{-T/2}^{T/2} s(t)\cos n\omega_1 t\,dt \\ b_n &= \frac{2}{T}\int_{-T/2}^{T/2} s(t)\sin n\omega_1 t\,dt \end{aligned} \tag{2.6}$$

Harmonics

Thus, in the general case a periodic signal contains a time-independent *constant component* and an infinite set of *harmonics* at frequencies $\omega_n = n\omega_1$, $n = 1, 2, 3, \ldots$, which are multiples of the fundamental frequency of the sequence.

Any harmonic in a Fourier series is characterized by its amplitude A_n and its initial phase φ_n. To show this, the expansion coefficients should be written as follows:

$$a_n = A_n \cos \varphi_n$$

$$b_n = A_n \sin \varphi_n$$

so that

$$A_n = \sqrt{a_n^2 + b_n^2}$$

and

$$\varphi_n = \arctan(b_n/a_n)$$

Substituting the above expressions into Eq. (2.5) yields an

equivalent form of the Fourier series

$$s(t) = a_0/2 + \sum_{n=1}^{\infty} A_n \cos(n\omega_1 t - \varphi_n) \tag{2.7}$$

which is sometimes more convenient to use.

The spectral diagram of a periodic signal. This term applies to a plot of the Fourier coefficients for a specific signal. It is usual to distinguish an amplitude diagram and a phase diagram (Fig. 2.1). The numbers laid off as abscissae are the discrete harmonic frequencies.

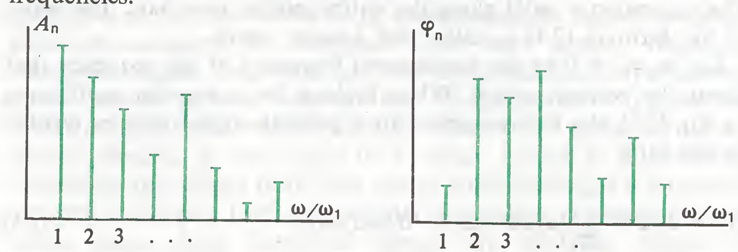


Fig. 2.1 Spectral diagrams of a periodic signal: (*a*) amplitude; (*b*) phase

Solve Problems 1 and 2

Of the two, the amplitude diagram is more frequently used as it contains information from which we can determine the percentage of a particular harmonic in the spectrum of the entire periodic signal.

Example 2.1. *The Fourier series of a periodic train of rectangular video pulses of known parameters* (τ, T, A), *which is even about the point* $t = 0$.

The pulse duty factor

In communication engineering, use is made of the ratio

$$q = T/\tau$$

which may be called the *reciprocal pulse duty factor**. Using Eqs. (2.6), we get

$$a_0/2 = A/q$$

$$a_n = \frac{2A}{T}\int_{-\tau/2}^{\tau/2} \cos n\omega_1 t\, dt = \frac{2A}{\pi n}\sin(n\omega_1 \tau/2)$$

This leads us to a Fourier series of the form

$$s(t) = (A/q)\left[1 + 2\sum_{n=1}^{\infty} \frac{\sin(n\pi/q)}{n\pi/q}\cos n\omega_1 t\right] \tag{2.8}$$

---

* In the UK and US literature on the subject, use is made of τ/T, called *the pulse duty factor.– Translator's note.*

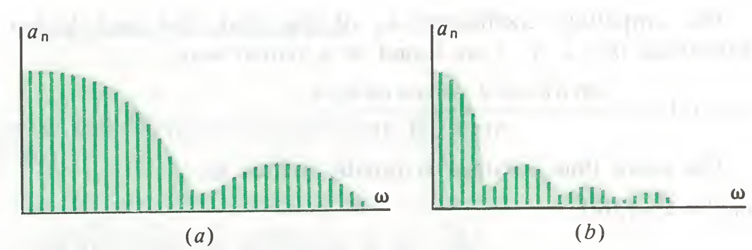


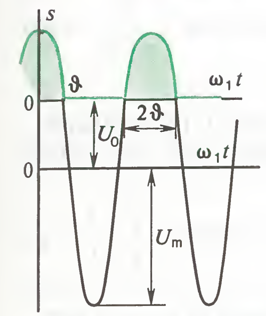
Fig. 2.2 Amplitude spectrum of a periodic sequence of rectangular video pulses: (*a*) for a large pulse period to pulse duration ratio; (*b*) for a small pulse period to pulse duration ratio

The typical amplitude diagrams of the sequence in question for two limiting cases are shown in Fig. 2.2.

It is important to note that a sequence of short pulses recurring at widely spaced intervals ($q \gg 1$) has a spectrum rich in harmonics.

The spectral diagram of the above form is said to have a *lobed structure.*

Example 2.2. *The Fourier series of a periodic pulse sequence formed by a harmonic signal of the form* $U_m \cos\omega_1 t$ *limited at* U_0 (*it is assumed that* $|U_0| < U_m$).



The cut-off angle

A special parameter, the *cut-off angle*, ϑ, is used to describe such a signal. It can be deduced from the relation

$$U_m \cos\vartheta = U_0$$

Hence, $\vartheta = \arccos(U_0/U_m)$. Then the quantity 2ϑ will be the duration of one pulse in electrical degrees.*

Analytically, the pulse repeated in the sequence can be written as

$$s(t) = U_m \cos\omega_1 t - U_0, \quad -\vartheta < \omega_1 t < \vartheta$$

The constant term of the sequence is

$$a_0/2 = \frac{1}{T}\int_{-\vartheta/\omega_1}^{\vartheta/\omega_1} (U_m \cos\omega_1 t - U_0)\, dt$$

$$= \frac{1}{2\pi}\int_{-\vartheta}^{\vartheta} (U_m \cos\omega_1 t - U_0)\, d(\omega_1 t) = \frac{U_m}{\pi}(\sin\vartheta - \vartheta\cos\vartheta)$$

The amplitude coefficient of the fundamental harmonic is

$$a_1 = \frac{1}{2\pi}\int_{-\vartheta}^{\vartheta} (U_m \cos\omega_1 t - U_0)\cos\omega_1 t\, d(\omega_1 t)$$

$$= (U_m/\pi)(\vartheta - \sin\vartheta\cos\vartheta)$$

---

* In the UK literature on the subject this quantity is called the *angle of current flow.* In the US literature, it is known as the *operating angle.* – *Translator's note.*

The amplitude coefficients a_n of the 2nd, 3rd and higher harmonics ($n=2, 3, \ldots$) are found in a similar way:

$$a_n=(2U_m/\pi)\frac{\sin n\vartheta\cos\vartheta-n\cos n\vartheta\sin\vartheta}{n(n^2-1)}$$

The result thus obtained is usually written as

$$a_0/2=U_m\gamma_0(\vartheta)$$

$$\ldots\ldots\ldots$$

$$a_n=U_m\gamma_n(\vartheta)$$

where $\gamma_0, \gamma_1, \gamma_2, \ldots$ are Berg functions:

$$\gamma_0(\vartheta)=(1/\pi)(\sin\vartheta-\vartheta\cos\vartheta)$$

$$\gamma_1(\vartheta)=(1/\pi)(\vartheta-\sin\vartheta\cos\vartheta) \tag{2.9}$$

$$\ldots\ldots\ldots$$

$$\gamma_n(\vartheta)=(2/\pi)\frac{\sin n\vartheta\cos\vartheta-n\cos n\vartheta\sin\vartheta}{n(n^2-1)}$$

for $n=2, 3, \ldots$.

A. I. Berg (1893-1979), a leading Soviet scientist in the field of telecommunications

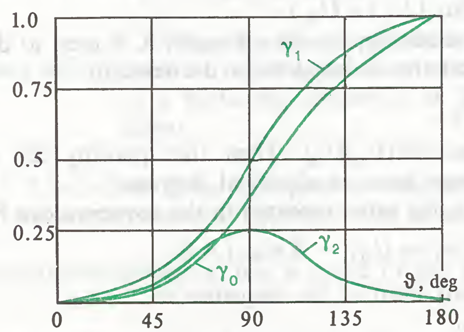


Fig. 2.3 Plots of the first few Berg functions

Plots of the functions γ_0, γ_1, and γ_2 are shown in Fig. 2.3. Since these functions are frequently used in engineering calculations, they are tabulated in an appendix to this book along with a computer subroutine in FORTRAN.

The complex form of the Fourier series. The principal formula (2.5) of Fourier analysis as applied to periodic signals can be given an elegant form, if we represent harmonic functions as a sum of exponentials with imaginary exponents. By applying Euler's formulas, we can re-write the series (2.5) as

$$s(t)=a_0/2+\sum_{n=1}^{\infty}\left(a_n\frac{e^{jn\omega_1 t}+e^{-jn\omega_1 t}}{2}+b_n\frac{e^{jn\omega_1 t}-e^{-jn\omega_1 t}}{2j}\right) \tag{2.10}$$

Instead of a_n and b_n, we introduce new coefficients

$$C_n=(a_n-jb_n)/2$$

for $n=1, 2, 3, \ldots$. We may as well determine the coefficients C_n for negative indices n, such that

$$C_{-n}=(a_n+jb_n)/2=C_n^*$$

because the coefficients a_n are even, and the coefficients b_n are odd with respect to the indices. Thus, the summation in (2.10) may be extended to include all values of n, both positive and negative:

$$s(t)=\sum_{n=-\infty}^{\infty}C_n\exp(jn\omega_1 t) \tag{2.11}$$

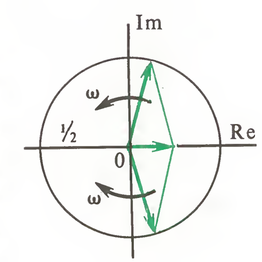
Equation (2.11) is the Fourier series in complex form. As can be readily seen, the complex coefficients are given by

$$C_n=\frac{1}{T}\int_{-T/2}^{T/2}s(t)\exp(-jn\omega_1 t)\,dt \tag{2.12}$$

The concept of negative frequencies. Since it is symmetrical about the origin of frequencies, the spectral diagram of a periodic signal in the form given by (2.11) contains components on the negative half-axis.

A few remarks are in order with regard to the concept of negative frequency. Consider the identity

$$\cos\omega_1 t=\frac{e^{j\omega_1 t}+e^{-j\omega_1 t}}{2}$$



In the method of complex amplitudes, the term $(1/2)\exp(j\omega_1 t)$ is represented on a complex plane as a phasor of length 1/2, which rotates at angular velocity ω_1 in the direction of increasing polar angle ωt. The phasor corresponding to the term $(1/2)\exp(-j\omega_1 t)$ differs in that it rotates in the opposite direction. On combining, these two complex numbers yield a real number.

In the series of Eq. (2.11), the terms with positive and negative frequencies form pairs. For example,

$$C_n\exp(jn\omega_1 t)+C_{-n}\exp(-jn\omega_1 t)=|C_n|\exp[j(n\omega_1 t+\varphi_n)]$$
$$+|C_n|\exp[-j(n\omega_1 t+\varphi_n)]=2|C_n|\cos(n\omega_1 t+\varphi_n)$$

To sum up, a negative frequency is not a physical, but

a mathematical concept associated with the manner in which complex numbers are represented.

Representation of a periodic signal on a complex plane. The structure of the Fourier series in Eq. (2.11) is such that a periodic signal can be represented by an infinite sum of phasors rotating in a complex plane (Fig. 2.4).

The construction involved is carried out as follows. From the origin (point 0), we draw a real phasor C_0 on the complex plane. This phasor represents the term whose number is $n=0$. Then, in Eq. (2.11) we set t equal to zero and construct a sum of phasors

$$C_+ = C_1 + C_2 + C_3 + \dots$$

$$C_- = C_{-1} + C_{-2} + C_{-3} + \dots$$

answering the contributions made by the positive-frequency and negative-frequency terms. For a converging Fourier series, each of the sums is represented by a phasor of finite length.

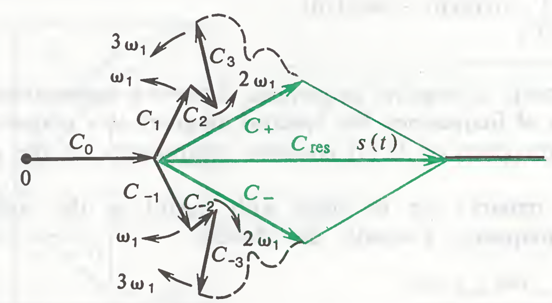


Fig. 2.4 Graphic representation of a Fourier series in complex form

As already noted, the coefficients of a Fourier series for positive and negative frequencies are complex conjugate. Therefore the $C_+ + C_-$ phasor is always real. When it is combined with the constant term C_0, it forms a phasor whose length, $s(0)$, is equal to the instantaneous value of the signal at the initial instant of time, t_0.

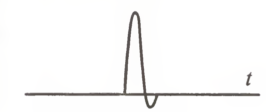
Then the picture is transformed: the phasors $C_1, C_2, \dots$ for positive frequencies rotate at angular velocities ω_1, $2\omega_1$, ... up in phase angle, whereas the phasors C_{-1}, C_{-2}, ... rotate in the opposite direction. The tip of the resultant phasor gives the current value of the signal.

The pictorial representation we have just examined is very helpful sometimes. We shall use it in the next section.

2.2 Spectral Analysis of Nonperiodic Signals. The Fourier Transform

Fourier analysis permits an in-depth and fruitful generalization with which we can readily obtain spectral characteristics for nonperiodic signals. Those of special interest to communication engineering are signals in the form of single pulses.

A single pulse



Periodic continuation of a pulse. Let $s(t)$ be a pulse signal of finite duration. If we mentally add to it similar signals recurring with period T, we shall obtain the already familiar periodic sequence, $s_{per}(t)$, which can be represented as a complex Fourier series

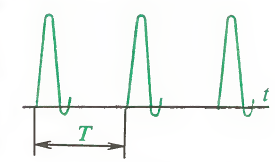
$$s_{per}(t) = \sum_{n=-\infty}^{\infty} C_n \exp(jn\omega_1 t) \tag{2.13}$$

with the complex coefficients given by the integrals

$$C_n = \frac{1}{T} \int_{-T/2}^{T/2} s_{per}(t) \exp(-jn\omega_1 t)\, dt \tag{2.14}$$

In order to go back to a single pulse signal, we should let the period T tend to infinity. Then,

A periodic sequence



(1) The frequencies of adjacent harmonics, $n\omega_1$ and $(n+1)\omega_1$, will be as close as we may wish, so that in Eqs. (2.13) and (2.14) we may replace the discrete variable $n\omega_1$ with a continuous variable ω, which is the current frequency.

(2) The amplitude coefficients C_n decrease without bound because the period T appears in the denominator of Eq. (2.14).

Our objective is to find the limiting form for Eq. (2.13) with T tending to infinity.

The concept of the spectrum of a signal. Let us take advantage of the fact that the coefficients of a Fourier series form complex conjugate pairs

$$C_n = A_n \exp(j\varphi_n)$$

$$C_{-n} = A_n \exp(-j\varphi_n)$$

Each pair represents a harmonic oscillation

$$A_n \exp[j(n\omega_1 t + \varphi_n)] + A_n \exp[-j(n\omega_1 t + \varphi_n)] = 2A_n \cos(n\omega_1 t + \varphi_n)$$

with a complex amplitude

$$2A_n \exp(j\varphi_n) = 2C_n$$

Consider a small interval of physical frequencies, $\Delta\omega$, in the

neighbourhood of some frequency ω_0. The number of spectral components contained within that interval is given by

$$N = \Delta\omega/\omega_1 = \Delta\omega T/2\pi$$

In physics, it is usually said that signals are combined coherently

Since their frequencies differ by an amount which may be made as small as we like, we may combine them as if they were of the same frequency and had the same complex amplitude

$$2C_n = \frac{2}{T}\int_{-\infty}^{\infty} s(t)\exp(-j\omega_0 t)\,dt$$

As a result, we obtain the complex amplitude of an equivalent harmonic signal representing the contributions of all the spectral components contained within the interval $\Delta\omega$:

$$\Delta A_{\omega_0} = \frac{2N}{T}\int_{-\infty}^{\infty} s(t)\exp(-j\omega_0 t)\,dt =$$

$$= \frac{\Delta\omega}{\pi}\int_{-\infty}^{\infty} s(t)\exp(-j\omega_0 t)\,dt \qquad (2.15)$$

Here we have defined a new function

$$S(\omega) = \int_{-\infty}^{\infty} s(t)\exp(-j\omega t)\,dt \qquad (2.16)$$

which is variously called the *amplitude spectral density* [20], the *complex spectral density**, the *amplitude spectrum* [20], or simply the *spectrum*** of the signal $s(t)$. It is the *Fourier transform of* the signal.

The physical significance of the spectrum. The result can be best interpreted if we change from the angular frequency ω to the cyclic frequency

$$f = \omega/2\pi$$

Then Eq. (2.15) will take the form

$$\Delta A_{f_0} = 2S(2\pi f_0)\Delta f \qquad (2.17)$$

It should be interpreted as follows: The spectrum $S(2\pi f_0) = S(\omega_0)$ is a scale factor connecting the length of the frequency interval Δf and the corresponding complex amplitude ΔA_{f_0} of the harmonic signal at the central frequency f_0.

A fact of fundamental importance is that the spectrum is

* S. Stein and J. Jones. *Modern Communication Principles.*

** C. W. Helstrom. *Statistical Theory of Signal Detection* Oxford: Pergamon Press, 1968. A. Peled and B. Liu. *Digital Signal Processing.* John Wiley and Sons, New York, 1976.– *Translator's note.*

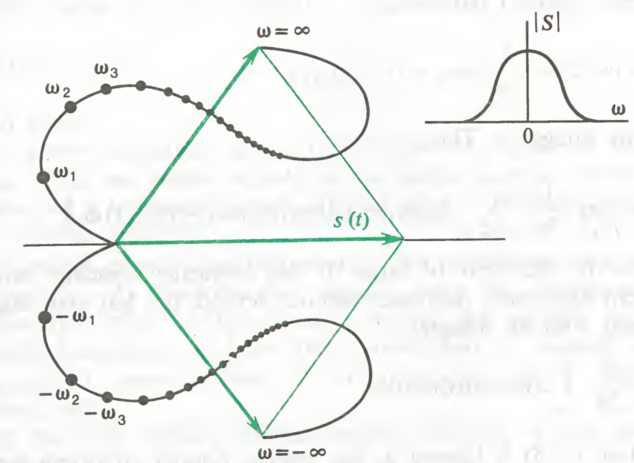


Fig. 2.5 Vector diagram of a nonperiodic signal (the plot on the right of the figure relates the magnitude of the spectral density to frequency)

a complex function of frequency and, as such, it carries information about both the amplitude and the phase of the elementary sine waves. On the phasor diagram of a nonperiodic signal (Fig. 2.5), the lengths of the elementary phasors are infinitesimal, so instead of the broken lines (T is finite) we obtain smooth curves (T tends to infinity). If we choose a set of equidistant points $0 < \omega_1 < \omega_2 < \dots$ on the frequency axis, the magnitude of the spectrum, $|S(\omega)|$, will establish a linear scale along the curves: an increase in the magnitude of the spectrum within the specified frequency range will lead to an increase in the spacing between the frequency points on the phasor diagram as well.

The signal waveform depends on both the magnitude and phase of the spectrum

This diagram has been plotted for some fixed instant of time. With time the shape of the curves changes in a very complex manner because the angular velocity at which the corresponding portions of the curves rotate increases with increasing frequency. Actually, however, what is important is not the shape of the curve, but the projection of its terminal point on the horizontal axis.

The inverse Fourier transform. Now we shall solve the inverse problem of the spectral theory of signals: We shall recover a signal from its spectrum which is deemed to be known.

Let us assume again that the nonperiodic signal is derived from a periodic sequence when its period tends to infinity. Using Eqs. (2.13) and (2.14), we may write

$$s(t) = \lim_{T\to\infty}\sum_{n=-\infty}^{\infty}\frac{1}{T}S(n\omega_1)\exp(jn\omega_1 t)$$

The coefficient $1/T$ is proportional to the difference in frequency

between adjacent harmonics:

$$1/T = \omega_1/2\pi = \frac{1}{2\pi}[n\omega_1 - (n-1)\omega_1]$$

for any integer n. Thus,

$$s(t) = \lim_{T\to\infty} \frac{1}{2\pi} \sum_{n=-\infty}^{\infty} S(n\omega_1)\exp(jn\omega_1 t)[n\omega_1 - (n-1)\omega_1]$$

Since in the limit of large T the frequency spacing between adjacent harmonics decreases without bound, the last sum may be replaced with an integral

$$s(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} S(\omega)\exp(j\omega t)\,d\omega \tag{2.18}$$

The inverse Fourier transform

Equation (2.18) is known as the *inverse Fourier transform* for the signal $s(t)$.

This brings us to a fundamental conclusion: *The signal $s(t)$ and its spectrum $S(\omega)$ are uniquely related by a Fourier transform pair:*

$$S(\omega) = \int_{-\infty}^{\infty} s(t)\exp(-j\omega t)\,dt$$
$$(2.19)$$
$$s(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} S(\omega)\exp(j\omega t)\,d\omega$$

Solve Problem 3

The technique of spectral expansion has substantially enriched signal theory. It often happens, for example, that the mathematical model of a signal in the form of the function $s(t)$, that is, in the time domain, is complicated and not easy to visualize. In contrast, the treatment of the same signal in the frequency domain, that is, with the aid of the function $S(\omega)$, proves very simple. What is more important, however, is that the spectral representation of signals offers a direct approach to the response analysis of a wide class of communication circuits and systems. These matters will be taken up in detail in Chapters 8 and 9.

The condition for the existence of the spectrum of a signal. In mathematics, it has been explored in detail what properties the function $s(t)$ must possess in order that its Fourier transform can exist.

Omitting proof and the rather subtle points that are involved in the very concept of the existence of a mathematical entity, we shall formulate the final result: For a signal $s(t)$ to have its spectrum $S(\omega)$, it is essential that the signal should be absolutely integrable,

and this means that the integral

$$\int_{-\infty}^{\infty} |s(t)|\,dt < +\infty$$

The absolute integrability of a signal

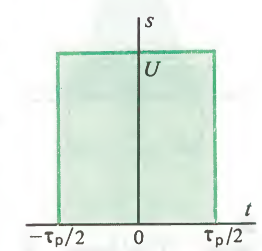
should exist.

The above condition substantially limits the class of eligible signals. Thus, we cannot speak, in the classical sense, about the spectrum of a harmonic signal, $s(t) = A\cos\omega_0 t$, defined along the entire infinite time axis.

Fortunately, present-day mathematics has techniques with which the spectra of nonintegrable signals can be evaluated in a reasonable way. In this case, however, the spectra will be generalized functions, and not the conventional or classical ones. The spectral representation of nonintegrable signals will be discussed later.

Now we shall consider several specific examples of how spectra are found.

The spectrum of a rectangular video pulse. Let there be a signal, $s(t)$, of amplitude U and duration τ_p, which is symmetrical about the origin of time. On the basis of Eq. (2.16),



$$S(\omega) = U\int_{-\tau_p/2}^{\tau_p/2} \exp(-j\omega t)\,dt = U\int_{-\tau_p/2}^{\tau_p/2}(\cos\omega t - j\sin\omega t)\,dt$$
$$= 2U\int_{0}^{\tau_p/2}\cos\omega t\,dt = (2U/\omega)\sin(\omega\tau_p/2)$$

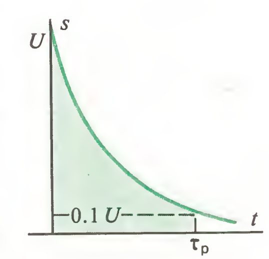
Thus, the spectrum of the signal in question is a real function of frequency. It is convenient to introduce a dimensionless variable, $\xi = \omega\tau_p/2$. Then the result can finally be written as follows:

$$S(\omega) = U\tau_p(\sin\xi/\xi) \tag{2.20}$$

It is interesting and important to note that at the zeroth frequency the spectrum is equal to the area of the pulse

$$S(0) = U\tau_p$$

A plot of the function in (2.20) is shown in Fig. 2.6.



The spectrum of an exponential video pulse. Consider the signal described by the function

$$s(t) = U\exp(-\alpha t)\sigma(t)$$

for the positive real value of the parameter α.

Strictly speaking, such a signal may only arbitrarily be called a pulse because of its behaviour at $t\to\infty$. However, the condition

between adjacent harmonics:

$$1/T = \omega_1/2\pi = \frac{1}{2\pi}[n\omega_1 - (n-1)\omega_1]$$

for any integer n. Thus,

$$s(t) = \lim_{T\to\infty} \frac{1}{2\pi} \sum_{n=-\infty}^{\infty} S(n\omega_1)\exp(jn\omega_1 t)[n\omega_1 - (n-1)\omega_1]$$

Since in the limit of large T the frequency spacing between adjacent harmonics decreases without bound, the last sum may be replaced with an integral

The inverse Fourier transform

$$s(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} S(\omega)\exp(j\omega t)\,d\omega \tag{2.18}$$

Equation (2.18) is known as the *inverse Fourier transform* for the signal $s(t)$.

This brings us to a fundamental conclusion: *The signal $s(t)$ and its spectrum $S(\omega)$ are uniquely related by a Fourier transform pair:*

$$S(\omega) = \int_{-\infty}^{\infty} s(t)\exp(-j\omega t)\,dt$$

$$s(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} S(\omega)\exp(j\omega t)\,d\omega \tag{2.19}$$

Solve Problem 3

The technique of spectral expansion has substantially enriched signal theory. It often happens, for example, that the mathematical model of a signal in the form of the function $s(t)$, that is, in the time domain, is complicated and not easy to visualize. In contrast, the treatment of the same signal in the frequency domain, that is, with the aid of the function $S(\omega)$, proves very simple. What is more important, however, is that the spectral representation of signals offers a direct approach to the response analysis of a wide class of communication circuits and systems. These matters will be taken up in detail in Chapters 8 and 9.

The condition for the existence of the spectrum of a signal. In mathematics, it has been explored in detail what properties the function $s(t)$ must possess in order that its Fourier transform can exist.

Omitting proof and the rather subtle points that are involved in the very concept of the existence of a mathematical entity, we shall formulate the final result: For a signal $s(t)$ to have its spectrum $S(\omega)$, it is essential that the signal should be absolutely integrable,

and this means that the integral

The absolute integrability of a signal

$$\int_{-\infty}^{\infty} |s(t)|\,dt < +\infty$$

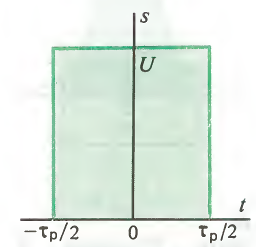
should exist.

The above condition substantially limits the class of eligible signals. Thus, we cannot speak, in the classical sense, about the spectrum of a harmonic signal, $s(t) = A\cos\omega_0 t$, defined along the entire infinite time axis.

Fortunately, present-day mathematics has techniques with which the spectra of nonintegrable signals can be evaluated in a reasonable way. In this case, however, the spectra will be generalized functions, and not the conventional or classical ones. The spectral representation of nonintegrable signals will be discussed later.

Now we shall consider several specific examples of how spectra are found.

The spectrum of a rectangular video pulse. Let there be a signal, $s(t)$, of amplitude U and duration τ_p, which is symmetrical about the origin of time. On the basis of Eq. (2.16),



$$S(\omega) = U \int_{-\tau_p/2}^{\tau_p/2} \exp(-j\omega t)\,dt = U \int_{-\tau_p/2}^{\tau_p/2} (\cos\omega t - j\sin\omega t)\,dt$$

$$= 2U \int_{0}^{\tau_p/2} \cos\omega t\,dt = (2U/\omega)\sin(\omega\tau_p/2)$$

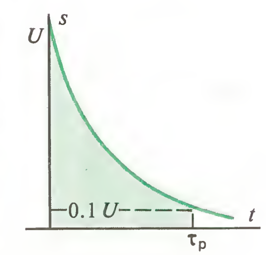
Thus, the spectrum of the signal in question is a real function of frequency. It is convenient to introduce a dimensionless variable, $\xi = \omega\tau_p/2$. Then the result can finally be written as follows:

$$S(\omega) = U\tau_p(\sin\xi/\xi) \tag{2.20}$$

It is interesting and important to note that at the zeroth frequency the spectrum is equal to the area of the pulse

$S(0) = U\tau_p$

A plot of the function in (2.20) is shown in Fig. 2.6.



The spectrum of an exponential video pulse. Consider the signal described by the function

$s(t) = U\exp(-\alpha t)\sigma(t)$

for the positive real value of the parameter α.

Strictly speaking, such a signal may only arbitrarily be called a pulse because of its behaviour at $t \to \infty$. However, the condition

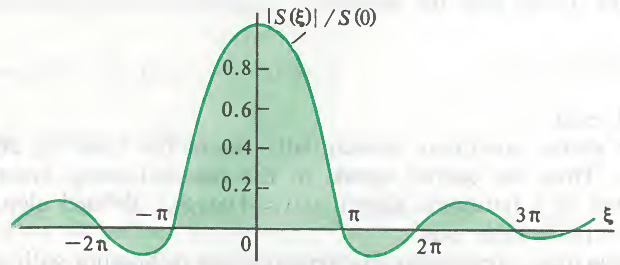


Fig. 2.6 Plot of the relative spectrum of a rectangular video pulse as a function of the parameter $\xi = \omega\tau_p/2$

$\alpha > 0$ assures a sufficiently rapid (exponential) decrease of the instantaneous values of the signal with increasing time. In communication engineering, the practical duration of such pulses is taken to be that over which the signal decreases to one-tenth of its original level:

$$\exp(-\alpha\tau_p) = 0.1$$

Hence,

$$\tau_p = 2.3026/\alpha$$

Solve Problems 5 and 6

The spectrum of an exponential video pulse is given by

$$S(\omega) = U\int_0^\infty \exp[-(\alpha + j\omega)t]\,dt$$
$$= [-U/(\alpha + j\omega)]\exp[-(\alpha + j\omega)t]\big|_{t=0}^{t=\infty}$$
$$= U/(\alpha + j\omega) \qquad (2.21)$$

The spectrum of an exponential waveform differs from that of a rectangular pulse in two fundamental aspects:

1. In accordance with Eq. (2.21), the spectrum $S(\omega)$ does not go to zero at any finite value of frequency.
2. The spectrum of an exponential pulse is a complex-valued function

$$S(\omega) = |S(\omega)|\exp[j\psi(\omega)]$$

with modulus $|S(\omega)| = U/\sqrt{\alpha^2 + \omega^2}$, and argument $\psi(\omega) = -\arctan(\omega/\alpha)$.

The respective plots appear in Figs. 2.7 and 2.8.

The spectrum of a Gaussian video pulse. This type of signal is described by a function of the form

$$s(t) = U\exp(-\beta t^2)$$

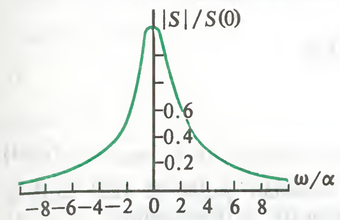


Fig. 2.7 The magnitude of the spectrum of an exponential pulse

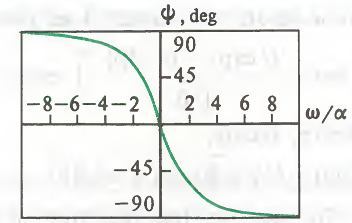
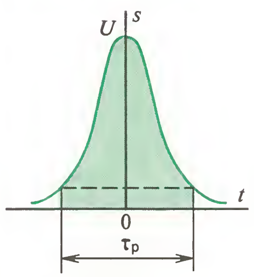


Fig. 2.8 The argument of the spectrum of an exponential pulse

This mathematical model is often used in cases where the pulse has a high degree of "smoothness". The effective duration of a Gaussian pulse is taken to be that over which the signal decreases to one-tenth of its maximum level. Referring to the plot, it can be seen that the duration τ_p must satisfy the relation



$$\exp[-\beta(\tau_p/2)^2] = 0.1$$

On re-arranging, we get

$$\tau_p = 2\sqrt{-\ln 0.1}/\sqrt{\beta} = 3.035/\sqrt{\beta} \qquad (2.22)$$

To determine the spectrum, we need to evaluate the integral

$$S(\omega) = U\int_{-\infty}^{\infty} \exp(-\beta t^2)\exp(-j\omega t)\,dt \qquad (2.23)$$

Now the integrand should be re-arranged in such a way that we may use the tabulated integral

$$\int_{-\infty}^{\infty} \exp(-x^2)\,dx = \sqrt{\pi}$$

For this purpose, from the exponent in (2.23) we isolate the complete square:

$$\beta t^2 + j\omega t = \beta t^2 + j\omega t - \omega^2/4\beta + \omega^2/4\beta$$
$$= (\sqrt{\beta}\,t + j\omega/2\sqrt{\beta})^2 + \omega^2/4\beta$$

Therefore,

$$S(\omega) = U\exp(-\omega^2/4\beta)\int_{-\infty}^{\infty}\exp\left[-\left(\sqrt{\beta}\,t + \frac{j\omega}{2\sqrt{\beta}}\right)^2\right]dt$$

Let us introduce a new variable

$$\xi = \sqrt{\beta}\,t + j\omega/2\sqrt{\beta}$$

such that

$$dt = d\xi/\sqrt{\beta}$$

As a result, the sought spectrum takes the form

$$S(\omega)=\frac{U\exp(-\omega^2/4\beta)}{\sqrt{\beta}}\int\limits_{-\infty}^{\infty}\exp(-\xi^2)d\xi$$

Hence, finally,

$$S(\omega)=U\sqrt{\pi/\beta}\exp(-\omega^2/4\beta) \tag{2.24}$$

To sum up, the spectrum of a Gaussian pulse is real and is described likewise by a Gaussian function of frequency.

The spectrum of the delta function. Let the signal $s(t)$ be a short pulse concentrated at point $t=0$ and having an area A. The mathematical model of this pulse is

$$s(t)=A\delta(t)$$

To find the spectrum of this signal, we need to evaluate the integral

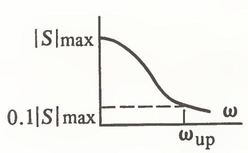
$$S(\omega)=A\int\limits_{-\infty}^{\infty}\exp(-j\omega t)\,\delta(t)\,dt$$

In Chapter 1 we have learned about the filtering properties of the delta function. This integral is equal to the value of the classical function at the point where the generalized function is concentrated. Therefore,

$$S(\omega)=A=\text{const} \tag{2.25}$$

This behaviour of the delta function spectrum is due to the initial idealization

Thus, the δ-impulse has a flat spectrum at all frequencies. It is interesting to interpret the result on a phasor diagram (see Fig. 2.5). Just as the impulse occurs ($t=0$), all elementary harmonic components which differ in frequency combine coherently, because in accord with Eq. (2.25) the spectrum is real. As the frequency is increased, their amplitudes do not decrease (compare with the previous examples). In consequence, the signal has an infinitely large magnitude at $t=0$. At all other times the phasor sum of the components is zero.



Relation between the duration and bandwidth of a pulse. From an analysis of the special cases examined so far, an important conclusion may be drawn: *As the duration of a pulse decreases, its bandwidth broadens.*

Bandwidth

Here and elsewhere, the bandwidth refers to the frequency interval within which the modulus of the spectrum is not smaller than some specified level. For example, it may range between $|S|_{max}$ and $0.1\,|S|_{max}$.

Consider a rectangular video pulse and assume that the upper frequency limit is the frequency at which the first zero occurs in the spectrum. This frequency can readily be found from the condition

$$\omega_{up}\tau_p/2=\pi$$

or

$$f_{up}\tau_p=1$$

In the case of an exponential video pulse, we arbitrarily assume that at the upper frequency limit the modulus of the spectrum reduces to one-tenth of its maximum value. Hence, it follows (see Eq. (2.20) and further) that

$$1/\sqrt{1+(\omega_{up}/\alpha)^2}=0.1$$

Solve Problem 4

or

$$\omega_{up}=\sqrt{99}\,\alpha$$

and

$$f_{up}=\omega_{up}/2\pi=1.584\alpha$$

Since the effective duration of an exponential pulse is

$$\tau_p=2.303/\alpha$$

the product

$$f_{up}\tau_p=3.647$$

Finally, the spectrum of the δ-impulse which has an infinitesimal duration extends along the frequency axis without bound.

To sum up, the product of the bandwidth by pulse duration is a constant number which depends solely on the pulse form and has, as a rule, a value of the order of unity:

$$f_{up}\tau_p=O(1)$$

It is said that the bandwidth and duration of a pulse are connected by the "uncertainty relation" (The term has been borrowed from quantum mechanics.)

The above relation is of primary importance for communication engineering. It sets the requirement for the bandwidth of a particular circuit or device when the signal duration is specified in advance. For example, as the pulse duration is reduced, the bandwidth of the amplifier that is to handle the signal must be increased in proportion. By the same token, short noise pulses which possess a broad bandwidth may impair the quality of reception over a significant frequency band.

2.3 Basic Properties of the Fourier Transform

In the preceding section we have learned how the Fourier transformation is applied and have found the spectra of simple, but frequently occurring pulse signals. Now we shall consider the properties of the Fourier transform.

Linearity of the Fourier transform. Given a set of signals $s_1(t)$, $s_2(t)$, ..., we can define the Fourier transform of their weighted sum by the formula

$$\sum_i r_i s_i(t) \leftrightarrow \sum_i r_i S_i(\omega) \tag{2.26}$$

Here r_i are arbitrary numerical coefficients. The validity of Eq. (2.26) can be proved by directly substituting the sums of signals in the Fourier transform (2.16).

Properties of the real and imaginary parts of the spectrum. Let $s(t)$ be a signal which takes on real values. In the general case, it has a complex spectrum

$$S(\omega) = \int_{-\infty}^{\infty} s(t)\cos\omega t\,dt - j\int_{-\infty}^{\infty} s(t)\sin\omega t\,dt = A(\omega) - jB(\omega)$$

On substituting the above expression in the Fourier inversion formula, Eq. (2.18), we get

$$s(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} [A(\omega) - jB(\omega)](\cos\omega t + j\sin\omega t)\,d\omega$$

For the signal subjected to a direct and an inverse Fourier transformation to remain real, we must require that

The integral of an odd function between symmetric limits is always equal to zero

$$\int_{-\infty}^{\infty} A(\omega)\sin\omega t\,d\omega = 0$$

and

$$\int_{-\infty}^{\infty} B(\omega)\cos\omega t\,d\omega = 0$$

This will occur only if the following condition is satisfied: *The real part $A(\omega)$ of the spectrum of the signal is an even function of frequency and the imaginary part $B(\omega)$ is an odd function of frequency:*

$$A(\omega) = A(-\omega)\quad B(\omega) = -B(-\omega) \tag{2.27}$$

The spectrum of a time-shifted signal. Suppose that for a signal $s(t)$ we know that

$$s(t) \leftrightarrow S(\omega)$$

Consider a similar signal, but occurring t_0 seconds later. Taking the point t_0 as the new origin of time, we can designate the translated signal as $s(t-t_0)$. Let us show that

$$s(t-t_0) \leftrightarrow S(\omega)\exp(-j\omega t_0) \tag{2.28}$$

This is so because

A change of variable $t - t_0 = x$

$$s(t-t_0) \leftrightarrow \int_{-\infty}^{\infty} s(t-t_0)\exp(-j\omega t)\,dt$$

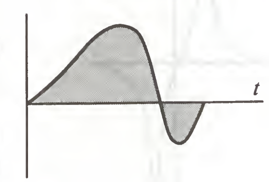
$$= \int_{-\infty}^{\infty} s(x)\exp(-j\omega t_0)\exp(-j\omega x)\,dx$$

$$= S(\omega)\exp(-j\omega t_0)$$

At any value of t_0 the complex number $\exp(-j\omega t_0)$ has a modulus of unity. Therefore, the amplitudes of the elementary harmonics that constitute the signal are independent of its position on the time axis. Information about the signal is embedded in the phase angle of its spectrum.

The dependence of the spectrum on the time scale adopted. Suppose that the original signal $s(t)$ is subjected to a transformation involving a change in the time scale. This implies that the role of time t is now played by the new independent variable kt (here k is some real number). If $k > 1$, the original signal is compressed in time. If $0 < k < 1$, the signal is expanded. It has been found that if $s(t) \leftrightarrow S(\omega)$, then

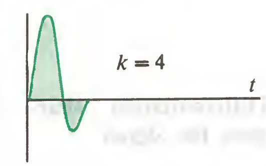
The original signal



$$s(kt) \leftrightarrow \frac{1}{k}S(\omega/k) \tag{2.29}$$

To demonstrate,

$$s(kt) \leftrightarrow \int_{-\infty}^{\infty} s(kt)\exp(-j\omega t)\,dt = \frac{1}{k}\int_{-\infty}^{\infty} s(x)\exp(-j\omega x/k)\,dx$$



The compressed signal

from which Eq. (2.29) follows directly.

Thus, if we wish to compress a signal in time while retaining its form, we should distribute the same spectral components over a broader frequency interval and reduce their amplitudes in proportion.

An interesting special case of time scale transformation is the reversal in time, $k = -1$. Using Eq. (2.29), we find the spectrum of the time-inverted signal:

$$s(t_{\text{inv}}) \leftrightarrow -S(-\omega) \tag{2.30}$$

In the case of time reversal, the modulus of the spectrum remains unchanged, but the positive and negative frequency regions of the spectrum are interchanged, and the initial phases of all harmonics are shifted through 180°.

Convolution

The integral on the right-hand side of Eq. (2.35) is called the *convolution of the functions* V and U. Symbolically, it will be designated as

$$\int_{-\infty}^{\infty} V(\xi)U(\omega-\xi)\,d\xi = V(\omega)*U(\omega)$$

Thus the spectrum of the product of two signals is equal, to within a constant, to the convolution of the spectra of the cofactors:

$$u(t)v(t) \leftrightarrow \frac{1}{2\pi} V(\omega)*U(\omega) \tag{2.36}$$

The reader can readily prove that the operation of convolution is commutative. This means that the order of the functions involved may be reversed:

$V(\omega)*U(\omega) = U(\omega)*V(\omega)$

Work Problems 7 and 8

The convolution theorem we have just proved may be reversed. If the spectrum of a signal can be represented as the product

$S(\omega) = S_1(\omega)S_2(\omega)$

such that $S_1(\omega) \leftrightarrow s_1(t)$ and $S_2(\omega) \leftrightarrow s_2(t)$, then the signal $s(t) \leftrightarrow S(\omega)$ is the convolution of the signals $s_1(t)$ and $s_2(t)$, but in the time rather than in the frequency domain:

$$S_1(\omega)S_2(\omega) \leftrightarrow \int_{-\infty}^{\infty} s_1(t-\xi)s_2(\xi)\,d\xi \tag{2.37}$$

We leave it for the reader to prove this formula.

2.4 Spectra of Nonintegrable Signals

Many of the mathematical models widely used in communication theory do not satisfy the requirement of absolute integrability. This implies that the Fourier transformation in its usual form is not applicable. Yet, as already noted, we may speak of the spectra of such signals, if we assume that these spectra may be described by generalized functions.

The spectrum of a time-invariant signal. The simplest nonintegrable signal is a constant quantity, $U_0 = \text{const}$. Suppose that this signal can be represented by an inverse Fourier transform

$$U_0 = \frac{1}{2\pi}\int_{-\infty}^{\infty} S(\omega)\exp(j\omega t)\,d\omega$$

in which the spectrum $S(\omega)$ is yet unknown. If we recall the filtering properties of the delta function, it is an easy matter to conclude that for this equality to be satisfied identically, we must set

$S(\omega) = 2\pi U_0\delta(\omega)$

Then,

$$U_0 \leftrightarrow 2\pi U_0\delta(\omega) \tag{2.38}$$

The physical meaning of the above result is easy to grasp: A time-invariant signal has a spectral component solely at the zeroth frequency.

The spectrum of a complex exponential signal. Let

$s(t) = \exp(j\omega_0 t)$

be a complex exponential signal of a known frequency ω_0. It is easy to see that this signal does not possess the property of absolute integrability, because the function $s(t)$ does not tend to any limit as t goes to $\pm\infty$.

Let us take the Fourier transform of the signal:

$$\exp(j\omega_0 t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} S(\omega)\exp(j\omega t)\,d\omega$$

and match the spectrum so that the equality turns to an identity. Using the filtering properties of the delta function, we immediately obtain an important result:

$$\exp(j\omega_0 t) \leftrightarrow 2\pi\delta(\omega-\omega_0) \tag{2.39}$$

The following points deserve to be noted.

(1) The spectrum of a complex exponential signal is zero everywhere except the point $\omega = \omega_0$, where it has a delta-singularity.

(2) The spectrum of a given signal is not symmetrical about the point $\omega = 0$, but is concentrated in the region of either positive or negative frequencies.

The spectrum of harmonic oscillations. Let

$s(t) = \cos\omega_0 t$

By Euler's formula,

$s(t) = [\exp(j\omega_0 t) + \exp(-j\omega_0 t)]/2$

From the spectrum of the complex exponential signal derived earlier and from the linearity of the Fourier transform, we may

immediately write the spectrum of a cosine signal as

$$\cos \omega_0 t \leftrightarrow \pi[\delta(\omega-\omega_0)+\delta(\omega+\omega_0)] \tag{2.40}$$

The reader himself can readily prove that for a sine signal the following relation is valid

$$\sin \omega_0 t \leftrightarrow -j\pi[\delta(\omega-\omega_0)-\delta(\omega+\omega_0)] \tag{2.41}$$

The two signals are real functions of time, therefore their spectra are even or odd functions of frequency.

The spectrum of an arbitrary periodic signal. Earlier, we analysed periodic signals with the aid of Fourier series. Now we can expand our idea about their spectral properties and describe periodic signals by use of Fourier transforms.

Let

$$s(t)=\sum_{n=-\infty}^{\infty} C_n \exp(jn\omega_1 t)$$

be a periodic signal defined by a complex Fourier series of its own. On the basis of Eq. (2.39) and recalling the linearity of Fourier transformation, we can immediately write the spectrum of this signal as

$$S(\omega)=2\pi \sum_{n=-\infty}^{\infty} C_n \delta(\omega-n\omega_1) \tag{2.42}$$

In configuration, the corresponding plot of the spectrum is identical to the usual spectral diagram of a periodic signal. The plot is formed by δ-impulses in the frequency interval $\pm n\omega_1$.

The spectrum of the switching function. As the last example of how to calculate the spectrum of a nonintegrable signal, we shall consider the spectrum of the switching function $\sigma(t)$ which, for simplicity, we shall define at all points except the point $t=0$ (cf. (1.2)):

$$\sigma(t)=\begin{cases}1, & t>0\\ 0, & t<0\end{cases}$$

To begin with, it should be noted that the switching function can be derived from an exponential video pulse by passing to the limit:

$$\sigma(t)=\begin{cases}\lim\limits_{\alpha\to 0}\exp(-\alpha t), & t>0\\ 0, & t<0\end{cases}$$

Therefore, in order to find the spectrum of the switching function, we should pass to the limit as $\alpha \to 0$:

The expression for the spectrum of an exponential pulse is used

$$\sigma(t)\leftrightarrow \lim_{\alpha\to 0}\frac{1}{\alpha+j\omega}$$

The direct passage to the limit, under which

$$\sigma(t)\leftrightarrow 1/j\omega$$

holds at all frequencies except at $\omega=0$. This case must be examined separately.

To begin with, we split the spectrum of an exponential signal into its real and imaginary parts:

$$\frac{1}{(\alpha+j\omega)}=\frac{\alpha}{\alpha^2+\omega^2}-\frac{j\omega}{\alpha^2+\omega^2}$$

The limiting value of the first term vanishes for any $\omega\neq 0$. On the other hand,

$$\int_{-\infty}^{\infty}\frac{\alpha\, d\omega}{\alpha^2+\omega^2}=\int_{-\infty}^{\infty}\frac{d(\omega/\alpha)}{1+(\omega/\alpha)^2}=\pi$$

irrespective of the magnitude of α. Hence

$$\lim_{\alpha\to 0}\frac{\alpha}{\alpha^2+\omega^2}=\pi\delta(\omega)$$

To sum up, the switching function and its spectrum are connected by a relation of the form

$$\sigma(t)\leftrightarrow \pi\delta(\omega)+1/j\omega \tag{2.43}$$

Inaccurately, Eq. (2.43) is sometimes written with only the second term

The presence of a delta-singularity in this spectrum at $\omega=0$ is an indication that this signal contains a constant term equal to 1/2.

The spectrum of radio pulses. As will be recalled, a radio pulse $u_r(t)$ can be defined as the product of a video pulse $u_v(t)$, which acts as the envelope, and a nonintegrable harmonic oscillation:

Solve Problem 16

$$u_r(t)=u_v(t)\cos(\omega_0 t+\varphi_0) \tag{2.44}$$

In finding the spectrum of a radio pulse, we assume to know the function $S_v(\omega)$, which is the spectrum of its envelope. The spectrum of a cosine signal with an arbitrary initial phase is obtained by generalizing Eq. (2.40):

$$\cos(\omega_0 t+\varphi_0)\leftrightarrow \pi[\delta(\omega-\omega_0)\exp(j\varphi_0) + \delta(\omega+\omega_0)\exp(-j\varphi_0)] \tag{2.45}$$

The spectrum of a radio pulse is the convolution of the spectra of two signals, $u_v(t)$ and $\cos(\omega_0 t + \varphi_0)$:

$$S_r(\omega) = \frac{1}{2}\int_{-\infty}^{\infty} S_v(\omega - \xi)\left[\delta(\xi - \omega_0)\exp(j\varphi_0) + \delta(\xi + \omega_0)\exp(-j\varphi_0)\right] d\xi$$

Recalling the filtering properties of the delta function, we obtain

$$S_r(\omega) = \frac{1}{2}\exp(j\varphi_0) S_v(\omega - \omega_0) + \frac{1}{2}\exp(-j\varphi_0) S_v(\omega + \omega_0) \tag{2.46}$$

The manner in which the spectrum of a video pulse is transformed when it is multiplied by an r.f. harmonic oscillation is illustrated in Fig. 2.9.

Thus, in spectral terms the change-over from a video pulse to a radio pulse implies *the translation of the spectrum of the video pulse into the r.f. range.* Instead of one maximum at $\omega = 0$, the spectrum displays two maxima at $\omega = \pm\omega_0$ which are halved in absolute value.

▲ **Solve Problem 10**

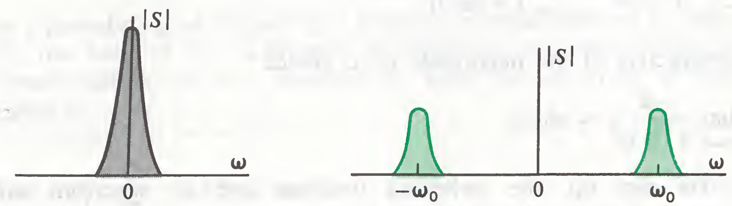


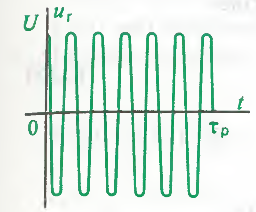
Fig. 2.9 Frequency dependence of the magnitude of the spectrum: (*a*) for a video pulse; (*b*) for a radio pulse

It is to be noted that the plots in Fig. 2.9 apply to the case where the frequency ω_0 is substantially higher than the effective bandwidth of the video pulse. (Precisely this case is usually realized in practice.) As is seen, there is no noticeable overlap between the spectra for positive and negative frequencies. However, it may so happen that the bandwidth of the video pulse is so broad (in the case of a short pulse) that the selected value of frequency ω_0 is not sufficient to avoid overlapping. In consequence, the spectra of the video pulse and of the radio pulse cease to be similar in shape.

Example 2.3. *Find the spectrum of a rectangular radio pulse.*

For simplicity, we set the initial phase in (2.44) equal to zero. Then the mathematical model of the radio pulse takes the form

$$u_r(t) = U[\sigma(t) - \sigma(t - \tau_p)]\cos\omega_0 t$$



Knowing the spectrum of the corresponding video pulse, Eq. (2.20), and using Eq. (2.46), we find the sought spectrum

$$S_r(\omega) = \frac{U\tau_p}{2}\left[\frac{\sin\dfrac{(\omega-\omega_0)\tau_p}{2}}{\dfrac{(\omega-\omega_0)\tau_p}{2}} + \frac{\sin\dfrac{(\omega+\omega_0)\tau_p}{2}}{\dfrac{(\omega+\omega_0)\tau_p}{2}}\right] \tag{2.47}$$

The results found by Eq. (2.47) for two typical cases are shown in Fig. 2.10. In the first case, the envelope encompasses 10 cycles of the r.f. carrier ($\omega_0\tau_p = 20\pi$), and the frequency ω_0 is sufficiently high to avoid any overlap. In the second case, there is only one cycle of the r.f. carrier ($\omega_0\tau_p = 2\pi$). Due to the superposition of the positive and negative frequency components, the spectrum of the radio pulse has the characteristic asymmetrical lobed structure.

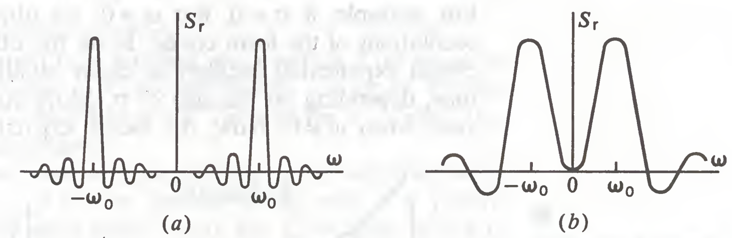


Fig. 2.10 Spectra of a radio pulse with a rectangular envelope: (*a*) for $\omega_0\tau_p = 20\pi$; (*b*) for $\omega_0\tau_p = 2\pi$

2.5 The Laplace Transform

In addition to Fourier transforms, communication theory widely uses another class of integral transforms to tackle a large variety of problems arising in the study of signals. They are Laplace transforms.

The concept of complex frequency. As we have seen, signal analysis by spectral methods is based on the fact that the signal of interest is represented as a sum of an infinitely large number of elementary components each of which varies in time as $\exp(j\omega t)$.

A natural generalization of the principle consists in that complex exponential signals with imaginary exponents are replaced with exponential signals of the form $\exp(pt)$, where p is a complex number such that

$$p = \sigma + j\omega$$

known as the *complex frequency.*

● **The complex frequency**

Any two of such complex signals can always be combined to

produce a real signal by, say, the following rule:

$$s(t) = \frac{1}{2}[\exp(pt) + \exp(p^* t)] \tag{2.48}$$

where

$$p^* = \sigma - j\omega$$

is the complex conjugate of p. To demonstrate,

$$s(t) = \exp(\sigma t)\frac{\exp(j\omega t) + \exp(-j\omega t)}{2} = \exp(\sigma t)\cos\omega t \tag{2.49}$$

A large variety of real signals can be derived, depending on the choice of the real and imaginary parts of the complex frequency. For example, if $\sigma = 0$, but $\omega \neq 0$, we obtain the usual harmonic oscillations of the form $\cos\omega t$. If, on the other hand, $\omega = 0$, we will obtain exponential oscillations either building up or decaying with time, depending on the sign of σ. More complex signal waveforms arise when $\omega \neq 0$. Now, the factor $\exp(\sigma t)$ plays the part of the

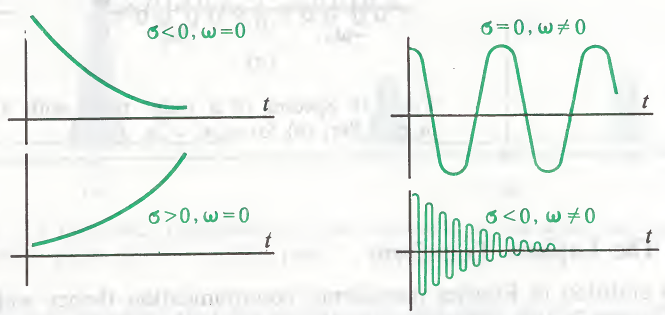


Fig. 2.11 Real signals corresponding to various values of the complex frequency

envelope exponentially varying with time. Some of the most typical signals are shown in Fig. 2.11.

The introduction of complex frequency is very fruitful above all because we can, without resorting to generalized functions, obtain the spectral representations of signals whose mathematical models are nonintegrable. More importantly, exponential signals of the form defined in Eq. (2.49) provide a "natural" means for studying signals in various linear systems. This matter will be taken up in detail in Chap. 8.

It is to be noted that the true physical frequency acts as the imaginary part of the complex frequency.

Basic relations. Let $f(t)$ be a signal, real or complex, defined at $t > 0$ and identically equal to zero at negative values of time. The Laplace transform, $F(p)$, of this signal is defined by the integral

$$F(p) = \int_0^\infty f(t)\exp(-pt)\,dt \tag{2.50}$$

The signal $f(t)$ is the *original time function*, and the function $F(p)$ is its *Laplace transform*.

The condition for the existence of the Laplace transform

The condition for the integral in (2.50) to exist consists in the following: The signal $f(t)$ may not grow faster than exponentially at $t > 0$, that is, it must satisfy the inequality

$$|f(t)| \leqslant k\exp(at) \tag{2.51}$$

where k and a are positive numbers.

When the above inequality is satisfied, the function $F(p)$ exists in the sense that the integral in (2.50) absolutely converges for all complex numbers p for which $\mathrm{Re}(p) > a$. The number a is called the *abscissa of absolute convergence*.

Relation between the Laplace and Fourier transforms

The variable p in the basic formula (2.50) may be identified with the complex frequency $p = \sigma + j\omega$. Indeed, in the case of a purely imaginary complex frequency when $\sigma = 0$, Eq. (2.50) goes into Eq. (2.16) defining the Fourier transform of a signal which vanishes at $t < 0$. Thus, the Laplace transformation should be looked upon as a generalization of the Fourier transformation to complex frequencies.

As with Fourier transforms, if we know a Laplace transform, we can recover the corresponding original time function. For this purpose, in the Fourier inversion formula

$$f(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} F(\omega)\exp(j\omega t)\,d\omega$$

we should perform an analytic continuation by changing from the imaginary variable $j\omega$ to the complex argument $(\sigma + j\omega)$. On the complex-frequency plane, it is usual to carry out integration along an infinitely extended vertical axis located to the right of the abscissa of absolute convergence. Since at $\sigma = \text{const}$ the differential $d\omega = (1/j)\,dp$, the inverse Laplace transform takes the form

$$f(t) = \frac{1}{2\pi j}\int_{c-j\infty}^{c+j\infty} F(p)\exp(pt)\,dp \tag{2.52}$$

In the theory of functions of a complex variable it is shown that Laplace transforms are well-behaved in terms of smoothness; they are analytic at all points on the complex p-plane, except the countable set of so-called singular points. The singular points are, as a rule, poles, single or repeated. Integrals of the form (2.52) can be evaluated by the techniques used in the theory of residues [11].

The Laplace transform and the inverse Laplace transform are referred to as a Laplace transform pair. In practice it is often convenient to look up each if the other is known in a set of tables. One such table is given in Appendix 4. It permits handling a sufficiently wide range of problems.

Examples of finding Laplace transforms. The calculation of Laplace transforms has much in common with finding Fourier transforms examined previously. Therefore, we shall limit ourselves to the most typical cases.

Example 2.4. *The Laplace transform of a generalized exponential pulse.*

Let

$$f(t) = \exp(p_0 t)\sigma(t)$$

where

$$p_0 = \sigma_0 + j\omega_0$$

is a fixed complex number. Owing to the presence of the σ-function, $f(t) = 0$ at $t < 0$. Using Eq. (2.50), we get

$$F(p) = \int_0^\infty \exp[-(p-p_0)t]\,dt = -\left.\frac{\exp[-(p-p_0)t]}{p-p_0}\right|_0^\infty$$

If $\operatorname{Re} p > \sigma_0$, then, on substituting the upper limit, the numerator vanishes, and we get the following correspondence

$$\exp(p_0 t)\sigma(t) \doteq 1/(p-p_0) \qquad (2.53)$$

As a special case of Eq. (2.53), we can obtain the Laplace transform of a real exponential video pulse:

$$\exp(-\alpha t)\sigma(t) \doteq 1/(p+\alpha) \qquad (2.54)$$

and of a complex exponential signal:

$$\exp(j\omega_0 t)\sigma(t) \doteq 1/(p-j\omega_0) \qquad (2.55)$$

Finally, on setting $\alpha = 0$ in Eq. (2.54), we obtain the Laplace

transform of the Heaviside function:

$$\sigma(t) \doteq 1/p \qquad (2.56)$$

Example 2.5. *The Laplace transform of the delta function.* If the δ-impulse occurs at time $t_0 > 0$, we need to evaluate the integral

$$\int_0^\infty \delta(t-t_0)\exp(-pt)\,dt = \exp(-pt_0)$$

Then the Laplace transform of the delta function takes the form

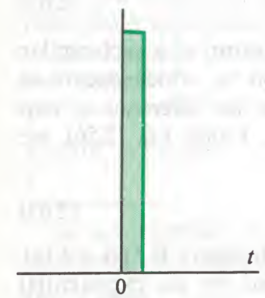
$$\delta(t-t_0) \doteq \exp(-pt_0) \qquad (2.57)$$

The Laplace transform in (2.57) is defined at all points on the complex p-plane. It has singularities nowhere except at infinity.

A certain difficulty may arise in taking the Laplace transform of the δ-impulse concentrated at $t = 0$, because it is not clear how we should treat the contribution from the generalized function concentrated at one of the ends of the range of integration. The point is that the delta function (see Ch. 1) is defined as the limit of a sequence of pulses, symmetrical about $t = 0$. If we proceed from formal considerations, then only a half of the pulse will fall within the range of integration, and this will result in halving the magnitude of the integral. To avoid this, the Laplace transform of $\delta(t)$ is defined as the limit

$$\lim_{\varepsilon \to 0} \int_{0-\varepsilon}^\infty \delta(t)\exp(-pt)\,dt = 1$$

independent of the parameter ε. With this approach, the delta function always belongs to the region of integration and has a Laplace transform of unit:



The δ-impulse belongs to the region $t > 0$

$$\delta(t) \doteq 1. \qquad (2.58)$$

2.6 Basic Properties of the Laplace Transform

Most properties of the Laplace transform are the same as the analogous properties of the Fourier transform. Therefore, in the subsequent discussion we will supply a proof only where the need to do so arises.

Linearity. The Laplace transformation is a linear integral transformation. Therefore, the weighted sum of signals is transform-

ed as follows:

$$\sum_i r_i f_i(t) \risingdotseq \sum r_i F_i(p) \qquad (2.59)$$

On the strength of this property, it is an easy matter to take the Laplace transforms of signals which may be represented as sums of relatively simple components for which the Laplace transforms are already known. For example, using Euler's formula and considering the Laplace transform pair in (2.55), we find that

$$\cos \omega_0 t\sigma(t) \risingdotseq p/(p^2+\omega_0^2) \qquad (2.60)$$

$$\sin \omega_0 t\sigma(t) \risingdotseq \omega_0/(p^2+\omega_0^2) \qquad (2.61)$$

The Laplace transform of a signal translated in time. If

$$f(t) \risingdotseq F(p)$$

we may write

$$f(t-t_0) \risingdotseq \exp(-pt_0)F(p) \qquad (2.62)$$

As an example, let us take the Laplace transform of a rectangular video pulse of unit amplitude and of duration τ_p, which occurs at $t=0$. It will suffice to note that this pulse is the difference of two switching functions translated in time by τ_p. Using Eq. (2.56), we obtain

$$f(t) \risingdotseq (1/p)[1-\exp(-p\tau_p)] \qquad (2.63)$$

The shifting theorem. Basically, it runs as follows: If $f(t) \risingdotseq F(p)$, the Laplace transform of a signal multiplied by an exponential function of time is obtained by shifting the argument of the Laplace transform:

$$f(t)\exp(-at) \risingdotseq F(p+a) \qquad (2.64)$$

The theorem is proved by the direct substitution of the function $f(t)\exp(-at)$ in (2.50).

By virtue of this property, we can, for example, take the Laplace transforms of exponential signals with a harmonic carrier. Thus, on the basis of (2.60) and (2.61), we get

$$\exp(-at)\cos\omega_0 t \risingdotseq \frac{p+a}{(p+a)^2+\omega_0^2} \qquad (2.65)$$

$$\exp(-at)\sin\omega_0 t \risingdotseq \frac{\omega_0}{(p+a)^2+\omega_0^2} \qquad (2.66)$$

The Laplace transform of the derivatives of a signal. The Laplace transform of the first derivative of a signal is taken by integration by parts

$$\frac{df}{dt} \risingdotseq \int_0^\infty (df/dt)\exp(-pt)\,dt$$

$$= f(t)\exp(-pt)\Big|_0^\infty + p\int_0^\infty f(t)\exp(-pt)\,dt$$

It is easy to see that the Laplace transform of the derivative contains the value of the signal at the initial point:

$$df/dt \risingdotseq pF(p)-f(0) \qquad (2.67)$$

The Laplace transform of the nth derivative is found by induction

$$d^n f/dt^n \risingdotseq p^n F(p) - p^{n-1}f(0) - p^{n-2}f'(0) - \ldots$$
$$- pf^{(n-2)}(0) - f^{(n-1)}(0) \qquad (2.68)$$

Since Laplace transforms contain the initial state of the signal at $t=0$, we may use the Laplace transformation to solve linear differential equations with known initial conditions [11].

The Laplace transform of an integral. If the signal is zero at $t=0$, then

$$\int_0^t f(\xi)\,d\xi \risingdotseq F(p)/p \qquad (2.69)$$

As an example of the above rule, let us take the Laplace transform of a ramp function:

$$f(t) \risingdotseq t\sigma(t)$$

To begin with, it is to be noted that $f(t)$ is the integral of the switching function

$$t\sigma(t) = \int_0^t \sigma(\xi)\,d\xi$$

Since

$$\sigma(t) \risingdotseq 1/p$$

it follows then that

$$t\sigma(t) \risingdotseq 1/p^2 \qquad (2.70)$$

The Laplace transform of the convolution of two signals. Similarly to the Fourier transform, the Laplace transform has the following property: There is a one-to-one correspondence between the convolution of two signals and the product of their Laplace transforms:

▲ **Solve Problem 11**

$$f_1(t)*f_2(t) \fallingdotseq F_1(p)F_2(p) \tag{2.71}$$

where

$$f_1(t)*f_2(t)=\int_0^t f_1(t-\xi)f_2(\xi)\,d\xi$$

The above relation provides a convenient means for taking the Laplace transform of a signal which may be *factored*, that is, represented by the product of two signals with known Laplace transforms.

Relationship between the limiting values of Laplace transforms and the original time functions. Let $F(p)$ be the Laplace transform of the time function $f(t)$. Then the following statement is valid: The behaviour of the Laplace transform near the origin of the coordinate system of the complex p-plane determines the nature of the original time function in the limit $t\to\infty$:

$$\lim_{p\to 0} pF(p)=\lim_{t\to\infty} f(t) \tag{2.72}$$

If the signal $f(t)$ does not contain delta singularities at zero, the formula

$$\lim_{p\to\infty} pF(p)=f(0) \tag{2.73}$$

also holds.

Summary

- The spectral representation of a signal is its decomposition into a sum (which may be finite or infinite) of elementary harmonic signals differing in frequency.
- A periodic signal can be represented by a Fourier series which is formed, generally speaking, by adding together an infinite number of harmonics at frequencies which are multiples of the fundamental repetition frequency of the signal sequence.
- The spectral representation of a nonperiodic (notably, pulse) signal is accomplished by expanding it into a Fourier integral.
- In the frequency domain, a signal is characterized by its spectrum.
- A signal and its spectrum are uniquely related by a Fourier transform pair.
- For the spectrum to exist in the classical sense, it is essential that the corresponding signal be absolutely integrable.
- The spectrum of a nonintegrable signal contains a singularity of the delta-function type.
- Exponential variations in the amplitude of oscillations with time are described in terms of complex frequency.

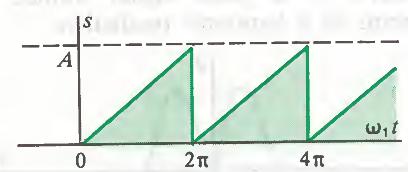
- The introduction of complex frequency in the Fourier transform results in a new type of integral transformation called the Laplace transform.
- Signals to which the Laplace transformation is applicable must be zero at $t<0$.

Review Questions

1. Why does the simple harmonic oscillation $\cos(\omega_0 t+\varphi_0)$ play an especially important role in communication engineering?
2. State the exact definition of a periodic signal. Name several physical processes for which the periodic signal model offers a sufficiently accurate representation.
3. Define the cut-off angle of a harmonic oscillation.
4. Define the concept of negative frequency. Explain its meaning.
5. What is the effect of coherent addition in the case of harmonic oscillations?
6. What properties should the spectrum of a real signal possess?
7. What is the practical limit for the duration of pulse signals?
8. What is a distinction of the spectrum of the δ-impulse?
9. How are the duration of a pulse and its bandwidth related?
10. How are the differentiation and integration of a signal represented in the frequency domain?
11. How are the spectra of a video pulse and of a radio pulse related?
12. What is the object of introducing the concept of complex frequency?
13. Describe the effect produced by the overlap of the frequency intervals in the spectrum of a radio pulse.
14. State the conditions that a signal should satisfy in order that its Laplace transform can be taken.
15. How is the Laplace transform of the product of two signals taken?
16. If we know the Laplace transform of a signal, how can we predict the asymptotic behaviour of the signal in the limit $t\to\infty$?

Problems

1. Show that the Fourier series of the sawtooth waveform shown in the figure



is given by the formula

$$s(t)=(A/2)-(A/\pi)[\sin\omega_1 t + (\sin 2\omega_1 t)/2 + (\sin 3\omega_1 t)/3+\ldots]$$

2. Find the amplitude coefficient of the 25th harmonic of a sawtooth signal, if $A=30$ V.

3. Show that if a periodic sequence is formed by the repetition of a pulse $s_0(t)$ of known spectrum $S_0(\omega)$, the complex amplitude of the nth member of the Fourier series is

$$C_n=(2/T)S_0(n\omega_1)$$

where T is the period of the sequence.

4. Given: a two-sided exponential video pulse

$$s(t)=U_0\exp(-\alpha|t|)$$

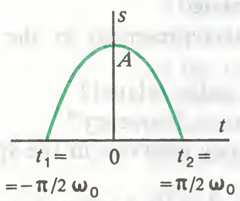
To find: An expression for its spectrum. Determine the duration and bandwidth of the signal.

5. Calculate the spectrum of the exponential video pulse defined in Eq. (2.21) with an amplitude of 20 V and a parameter $\alpha = 10^6\ s^{-1}$ at a frequency

$\omega_0 = 2 \times 10^5\ s^{-1}$ *

6. At what frequency will the spectrum of the pulse in Problem 5 have a phase angle of $-45°$?

7. Verify that the spectrum of the single cosine pulse shown in the accompanying figure



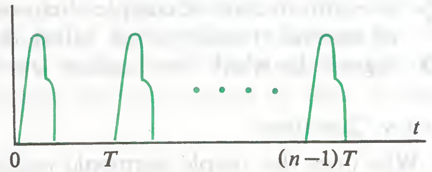
and defined as

$$s(t) = \begin{cases} 0, & t < t_1 \\ A\cos\omega_0 t, & t_1 < t < t_2 \\ 0, & t > t_2 \end{cases}$$

can be found by the formula

$$S(\omega) = A\left[\frac{\sin\dfrac{\pi(\omega_0+\omega)}{2\omega_0}}{\omega_0+\omega} + \frac{\sin\dfrac{\pi(\omega_0-\omega)}{2\omega_0}}{\omega_0-\omega}\right]$$

8. Find the spectrum of a train of n identical video pulses, as shown in the accompanying figure:



Show that the spectrum of the train is

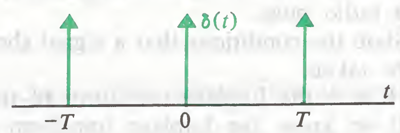
$$S_\Sigma(\omega) = S_0(\omega)\frac{1-\exp[-j(n+1)\omega T]}{1-\exp(-j\omega T)}$$

where $S_0(\omega)$ is the spectrum of a single pulse.

Hint: Use the formula for the summation of a geometrical progression

$$\sum_{i=0}^{n} r^i = \frac{1-r^{n+1}}{1-r}$$

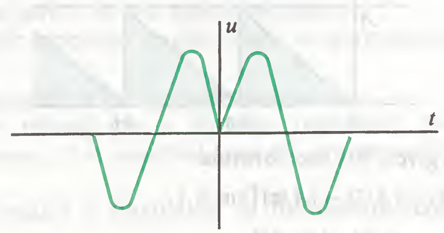
9. The train consists of three identical δ-impulses as shown in the figure



Show that the frequency dependence of the magnitude of the spectrum of the train is given by the formula

$$|S_\Sigma(\omega)| = [(1+\cos\omega T + \cos 2\omega T)^2 + (\sin\omega T + \sin 2\omega T)^2]^{1/2}$$

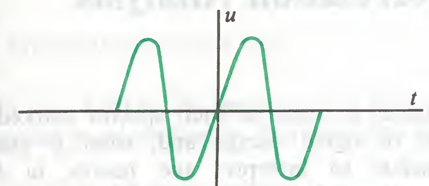
10. The accompanying figure shows the waveform of a pulse signal formed by segments of a harmonic oscillation.



Prove that the spectrum of the signal is zero both at the zeroth frequency and at the

* In this text the unit of angular frequency is taken to be second^{-1} [s^{-1}] – *Translator's note.*

carrier frequency. How will the spectrum of the signal change, if its waveform is



11. Using the convolution theorem, find the signal for which the Laplace transform is

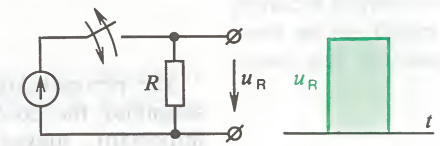
$F(p) = U_0/(p+\alpha)(p+\beta)$

Advanced Problems

12. Let there be a periodic signal described by a time function with jumps in level (discontinuities of the 1st kind). Show that, as their order increases, the Fourier coefficients of the signal have an asymptotic behavior of the form $O(1/n)$, irrespective of the form of the function.

13. Assuming the conditions of the previous problem, consider a signal in which the first derivative experiences discontinuities, whereas the function is continuous. Show that the Fourier coefficients have an asymptotic behaviour of the form $O(1/n^2)$.

14. Discuss the following "paradox": If we close for some time the switch in the circuit shown in the accompanying figure



a rectangular pulse $u_R(t)$ will develop across the load resistor. This pulse is a sum of the harmonic components existing at all times, including the time prior to the occurrence of the pulse. How does this agree with the assumption that no pulse may be produced although the harmonic components exist already?

15. Using the Fourier transform, find the integral representation for the derivatives of any order of the delta function.

16. Show that the spectrum of the σ-function (see Eq. (2.43)), when it is represented by the inverse Fourier transform, yields at time $t = 0$ a signal $\sigma(0)$ whose value is 1/2.

Hint: Take the frequency ω as a complex variable and evaluate the integral

$$\frac{1}{2\pi}\int_{-\infty}^{\infty}[\pi\delta(\omega) + 1/j\omega]\,d\omega$$

by the theory of residues.

Chapter 3

Power Spectra of Signals. Principles of Correlation Analysis

The representation of signals in terms of their spectra markedly simplifies the computation of signal energy and, what is more important, makes it possible to interpret the results in an easy-to-grasp form. These physical ideas are useful in various theoretical and applied fields of telecommunications. This chapter will examine the relations between the spectral and power characteristics of signals.

3.1 Cross-Spectral Density. Power Spectrum

In Chap. 1 we have introduced a fundamental characteristic of a system of two signals, $u(t)$ and $v(t)$, – their scalar product

$$(u,\ v)=\int_{-\infty}^{\infty} u(t)v(t)\,\mathrm{d}t \tag{3.1}$$

which is proportional to the cross energy of the signals. If the signals are identically the same, $u(t)=v(t)$, their scalar product changes to the signal energy

$$E_u=(u,\ u)=\int_{-\infty}^{\infty} u^2(t)\,\mathrm{d}t \tag{3.2}$$

Let us establish the relation between the scalar product of signals and their spectra.

The generalized Rayleigh formula. Assume that the two signals, $u(t)$ and $v(t)$, appearing in Eq. (3.1) are defined in terms of their spectra

$$u(t)=\frac{1}{2\pi}\int_{-\infty}^{\infty} S_u(\omega)\exp(j\omega t)\,\mathrm{d}\omega$$

$$v(t)=\frac{1}{2\pi}\int_{-\infty}^{\infty} S_v(\omega)\exp(j\omega t)\,\mathrm{d}\omega \tag{3.3}$$

On substituting $v(t)$ from (3.3) into (3.1) and reversing the order of integration with respect to time and frequency, we get

$$(u,\ v)=\frac{1}{2\pi}\int_{-\infty}^{\infty} \mathrm{d}\omega S_v(\omega)\int_{-\infty}^{\infty} u(t)\exp(j\omega t)\,\mathrm{d}t$$

Note that the inner integral in the last equation is the spectrum

of the signal $u(t)$ as found for the negative value of the argument:

$$\int_{-\infty}^{\infty} u(t)\exp(j\omega t)\,\mathrm{d}t=S_u(-\omega)$$

In our further discussion we shall assume that the signals in question are described by real functions of time. Then, as is easy to see,

$$(u,\ v)=\frac{1}{2\pi}\int_{-\infty}^{\infty} S_v(\omega)S_u^*(\omega)\,\mathrm{d}\omega \tag{3.4}$$

The above relation is called the *generalized Rayleigh formula*. It can be interpreted in a form which is easy to memorize: *The scalar product of two signals is proportional to the scalar product of their spectra*.

Of course, the order of complex conjugation in Eq. (3.4) may be reversed to give

$$(u,\ v)=\frac{1}{2\pi}\int_{-\infty}^{\infty} S_u(\omega)S_v^*(\omega)\,\mathrm{d}\omega \tag{3.5}$$

In the general case, the integrands in Eqs. (3.4) and (3.5) are complex, although the left-hand sides of the equalities are known to be real. The two equations can be re-arranged so that the real function of frequency appears under the integral. The point is that the integration is carried out between symmetrical limits and the integrands corresponding, to two symmetrical points, $\pm\omega$, are complex conjugates. To demonstrate,

$$S_u(-\omega)S_v^*(-\omega)=S_u^*(\omega)S_v(\omega)=[S_u(\omega)S_v^*(\omega)]^*$$

Since for any complex number z

$$z+z^*=2\operatorname{Re}z$$

we may introduce a real function

$$W_{uv}(\omega)=\operatorname{Re}[S_u(\omega)S_v^*(\omega)] \tag{3.6}$$

in terms of which the scalar product of the signals u and v may be written as follows:

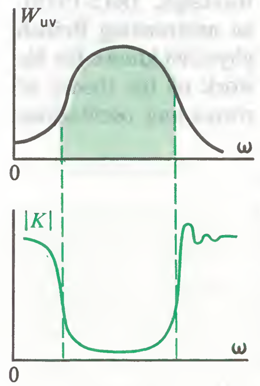
$$(u,\ v)=\frac{1}{2\pi}\int_{-\infty}^{\infty} W_{uv}(\omega)\,\mathrm{d}\omega \tag{3.7}$$

J. W. Strutt (Lord Rayleigh, 1842-1919), an outstanding British physicist known for his work on the theory of waves and oscillations

This is also known as Parseval's relation

Work Problem 1

The cross-power spectrum



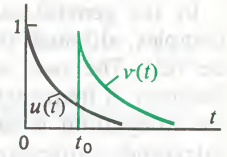
The amplitude response of an orthogonalizing filter as a function of frequency

The function $W_{uv}(\omega)$ is termed the *cross-power spectrum** of the signals u and v.

Equation (3.7) gives a deeper insight into the fine structure of the relation between the two signals. For one thing the various portions of their spectra contribute, in the general case, differently to the cross-power of the signals. The contribution is a maximum from the frequency intervals where the spectra overlap. For another, the generalized Rayleigh formula (3.7) suggests how the association between the signals can be minimized so that they become orthogonal in the limit. To achieve this, one of the signals must be matched to a special physical system called the *frequency filter*. This filter is to meet the following requirement: It may not pass those of the spectral components of the signal which lie within the frequency interval where the cross-power spectrum is a maximum. The frequency response of this orthogonalizing filter has a sharply defined minimum within that frequency interval.

The above approach to finding the scalar product, based on the concept of the cross-power spectrum, is directly related to the results obtained in Chap. 1, when we found the scalar product of signals expanded in terms of orthogonal basis elments. The difference is that now we use the continuous rather than discrete Fourier presentation.

Example 3.1. *The cross-power spectrum of two exponential video pulses of the same waveform, separated by a time interval* t_0.



Assuming that the two pulses have the same amplitude, their spectra may be written as follows:

$$u(t) = \exp(-\alpha t)\sigma(t)$$

$$\leftrightarrow S_u(\omega) = 1/(\alpha + j\omega)$$

$$v(t) = \exp[-\alpha(t - t_0)]\,\sigma(t - t_0) \leftrightarrow S_v(\omega)$$

$$= \exp(-j\omega t_0)/(\alpha + j\omega)$$

Then we find the product of the spectra:

$$S_u(\omega)S_v^*(\omega) = \exp(j\omega t_0)/(\alpha^2 + \omega^2)$$

* Some authors (e.g. A. Breipohl, Probabilistic Systems Analysis) call it the *cross-power density spectrum.* Others (Middleton [20]) call it the *cross-spectral density.– Translator's note.*

and, finally, the cross-power spectrum

$$W_{uv}(\omega) = \cos\omega t_0/(\alpha^2 + \omega^2) \qquad (3.8)$$

If we assign a fixed value to the parameter α which defines the waveform of the signals, the frequency behaviour of the cross-power spectrum will be wholly dependent on the time shift between the signals. Two typical plots for the function $W_{uv}(\omega)$ appear in Fig. 3.1.

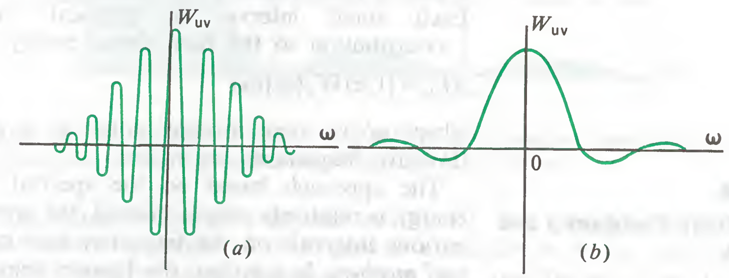
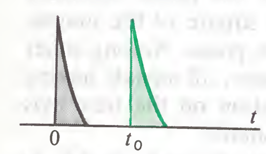


Fig. 3.1 Cross-power spectrum of two exponential video pulses: (a) for $\alpha t_0 \gg 1$; (b) for $\alpha t_0 \ll 1$



Of special interest is the case where the product αt_0 is small, that is, when the pulses substantially overlap in time. Equation (3.8) and the plot in Fig. 3.1*b* show that the cross-power spectrum has a well-defined low-frequency character. Hence we may conclude: In order to reduce the scalar product of such signals and to make them better resolvable, use should be made of a *high-pass filter* which suppresses all frequencies below some limiting, or cut-off, frequency, w_c.

Orthogonalization of pulses

Due to the action of a high-pass filter, the rapidly changing leading edge of the output pulse is produced by the high-frequency spectral components which are free to pass through. At the same time, the low-frequency components are filtered out and the duration of the output pulse is reduced substantially. As a result, the overlap of the pulses can be made as small as we like, and the pulses appearing at the output of the high-pass filter are orthogonal very nearly.

The power spectrum of a signal. The spectral representation of signal power can readily be developed as a special case of the generalized Rayleigh formula, if we deem the signals $u(t)$ and $v(t)$ to be the same. Then Eq. (3.6) defining the cross-power spectrum takes the form

$$W_u(\omega) = S_u(\omega)S_u^*(\omega) = |S_u(\omega)|^2 \qquad (3.9)$$

The power spectrum

The quantity $W_u(\omega)$ is called the *power spectrum* of a signal*. Then Eq. (3.7) may be re-written as

The Rayleigh formula

$$E_u = \int_{-\infty}^{\infty} u^2\, dt = \frac{1}{2\pi}\int_{-\infty}^{\infty} W_u(\omega)\, d\omega \tag{3.10}$$

The relation defined in Eq. (3.10) is known in various fields of physics as the (narrow-sense) *Rayleigh formula.* It states an important result: *The energy of any signal may be represented as a sum of contributions from various intervals on the frequency axis.* Each small interval of physical frequencies, $\Delta\omega$, makes a contribution to the total signal energy, equal to

$$\Delta E_u = (1/\pi)\, W_u(\omega')\, d\omega$$

where ω' is some interior point in a given frequency interval. (Positive frequencies are meant.)

Work Problems 3 and 4

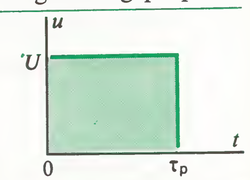
The approach based on the spectral representation of signal energy is relatively simple. Indeed, the energies associated with the various intervals on the frequency axis can be added together as *real numbers.* In contrast, the Fourier transform method as applied to the signals themselves is based on the fact that the complex amplitudes defining the contributions from small frequency intervals are added together as *complex numbers.*

When we analyse signals on the basis of their power spectra, we inevitably lose the information embedded in the phase spectrum because the power spectrum, Eq. (3.9), is the square of the magnitude of the spectrum and is independent of its phase. Among other things, when presented in terms of power spectra, all signals having the same waveform but differing in their position on the time axis appear completely identical and undistinguishable.

Yet, the concept of the power spectrum is very useful in estimating the real bandwidth of a signal for engineering purposes.

Example 3.2. *The power spectrum of a rectangular video pulse.* To obtain it, we should square the spectrum of (2.20):

$$W_u(\omega) = U^2\tau_p^2 \frac{\sin^2(\omega\tau_p/2)}{(\omega\tau_p/2)^2} \tag{3.11}$$



* This quantity is also called the *spectral density* (Helstrom, *Statistical Theory of Signal Detection.* Oxford: Pergamon Press, 1968), the *energy* (or *power*) *density* (D. Middleton, *An Introduction to Statistical Communication Theory.* New York, Toronto, London: McGraw Hill Book Co., Inc., 1960), the *power density spectrum* (*Handbook of Automation, Computation and Control* ed. by Eu. M. Grabbe, S. Ramo, and D. E. Wooldridge. New York: John Wiley and Sons, Inc., 1958), the *power spectral density* (W. C. Lindsey and M. K. Simon, *Telecommunication Systems Engineering,* Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1973), etc.–*Translator's note.*

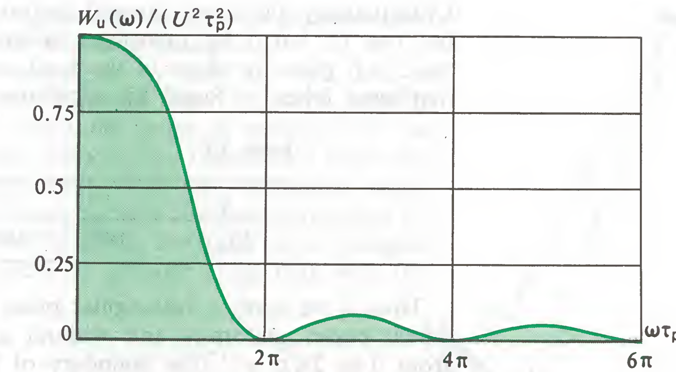


Fig. 3.2 The normalized power spectrum of a rectangular video pulse as a function of the nondimensional frequency variable $\omega\tau_p$

The corresponding plot appears in Fig. 3.2. It is clearly seen that the power spectrum of the signal is a maximum in the low-frequency region. As the frequency rises, the contributions from the various spectral components show a non-monotonic, oscillatory behaviour, but the general trend is the decrease in the power spectrum in inverse proportion to the square:

$W_u(\omega) = O(1/\omega^2)$ as ω tends to infinity

(rather than in inverse proportion to the first power, as is the case with the spectrum of the signal in question).

Formula (3.11) can be used to verify the Rayleigh formula. To begin with, we readily find the energy of the video pulse in the time domain (see Chap. 1):

$$E_u = U^2\tau_p \tag{3.12}$$

In order to find the energy in the frequency domain, we should evaluate the integral:

$$E_u = (U^2\tau_p^2/\pi)\int_0^{\infty} \frac{\sin^2(\omega\tau_p/2)}{\omega^2\tau_p^2/4}\, d\omega \tag{3.13}$$

A simple change of variable leads to Eq. (3.12) at once.

Energy distribution in the spectrum of a rectangular video pulse. It is interesting and, in many applied cases, important to know the share of the total energy within one, two, three, etc. lobes of the spectral diagram shown in Fig. 3.2. Let $E_{(k)}$ denote the energy of a rectangular video pulse, contained in a number k of consecutive lobes. By the Rayleigh formula,

$$E_{(k)} = (2/\pi)\, U^2\tau_p \int_0^{k\pi} \sin^2\xi/\xi^2\, d\xi \tag{3.14}$$

Unfortunately, the above integral cannot be evaluated analytically, but can be found by numerical integration. The accompanying table, 3-1, gives the share of the total energy in the consecutively numbered lobes, as found by calculation.

Table 3-1

k	1	2	3
$E_{(k)}/E$	0.902	0.950	0.967

Thus, if we apply a rectangular pulse to an ideal low-pass filter which passes uniformly and without attenuation all frequencies from 0 to $2\pi/\tau_p$ s^{-1} (the boundary of the first lobe), the output signal will account for 90.2% of the input energy.

As already noted, this approach to estimating the actual bandwidth of a signal does not give proper insight into the process. Among other things, we know nothing about the corruption that the signal suffers as it passes through. However, in cases where the shape of an oscillation is of secondary importance, and we are, above all, interested in the magnitude of signal energy (this situation is frequently encountered in statistical communication theory), the estimation of the bandwidth in terms of energy is especially warranted.

For example, from Table 3-1 it is seen that in going from $k = 1$ to $k = 2$ (which implies a two-fold increase in the bandwidth of the device through which the signal is passed) the energy of the wanted signal increases by a mere 4.8%. On the other hand, it is clear that noise, if there is any, may, as a result, double in energy if its power spectrum is flat over the frequency band of interest. That is why an unwarranted increase in bandwidth must be avoided.

▲ Solve Problem 5

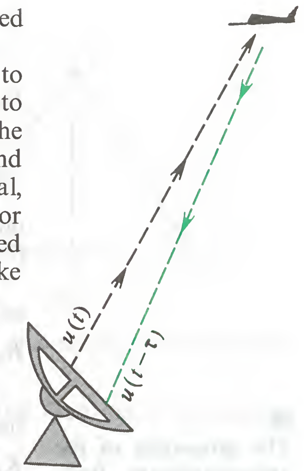
3.2 Correlation Analysis of Signals

In the early days of telecommunications, the choice of the best signals for specific applications was not critical. For one thing, the messages (telegraph signals, radio broadcasts) were simple in structure. For another, it was difficult to implement signals of complex waveforms and the associated equipment for their coding, modulation, demodulation, and decoding.

Today, the situation is entirely different. In present-day telecommunication complexes, the choice of signals is dictated not by the technical convenience in their generation, transformation and reception, but, above all, by the ability to achieve the goals envisaged at the design stage in an optimal manner, that is, with the utmost efficacy, using suitable signals. In order to get an idea

about how the need arises in signals possessing some predetermined properties, let us consider the following example.

Comparison of signals translated in time. Let us turn to a simplified description of how a pulsed radar measures range to a target. Here, information about the target is embedded in the magnitude of τ, the time delay between the transmitted pulse and the received echo. Characteristically, both the transmitted signal, $u(t)$, and the received signal, $u(t-\tau)$, have identical waveforms for any value of time delay. In block-diagram form, the device designed to process radar signals in order to measure range may look like the one shown in Fig. 3.3.



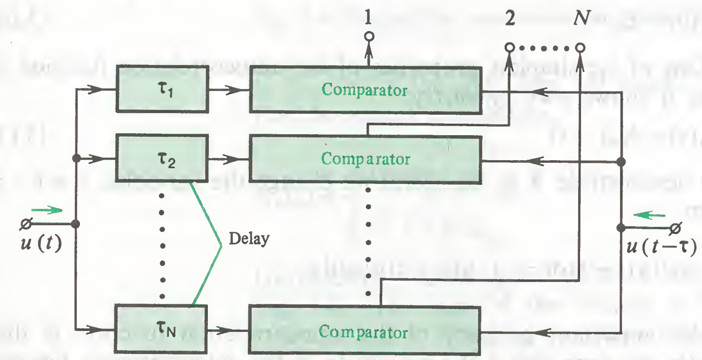


Fig. 3.3 Set-up for measuring delay time

The system consists of a set of elements which delay the transmitted "reference" signal for fixed time intervals $\tau_1, \tau_2, \ldots, \tau_N$. The delayed signals are applied together with the received echo signal to comparators which generate an output signal only if the two input signals are identical. If we know the No. of the channel in which this event occurs, we can measure the time delay and, in consequence, the range to the target.

The performance of the device improves progressively with the increase in difference between the transmitted signal and its echo translated in time. If the difference is negligible, an ambiguous indication may will occur, because signals will appear at the outputs of several adjacent comparators at the same time.

Thus, we have developed a qualitative idea about which signals may be taken as "well-behaved" for a given application.

Now we will pass on to the exact mathematical statement of the problem at hand and show that all these matters have a direct bearing on the power spectra of signals.

The autocorrelation function of a signal. In order to evaluate the degree of difference between a signal, $u(t)$, and its time-translated replica, $u(t-\tau)$, it is usual to resort to the *autocorrelation function*

$K_u(\tau)$ equal to the scalar product of the two signals:

$$K_u(\tau) = \int_{-\infty}^{\infty} u(t)\,u(t-\tau)\,dt \tag{3.15}$$

In our further discussion, it will be assumed that the signal of interest is a pulse localized in time, so the integral of the form in (3.15) does exist.

It follows directly from Eq. (3.15) that at $\tau = 0$ the autocorrelation function becomes equal to the signal energy:

$$K_u(0) = E_u \tag{3.16}$$

The properties of the autocorrelation function

One of the simplest properties of the autocorrelation function is that it shows *even symmetry*:

$$K_u(\tau) = K_u(-\tau) \tag{3.17}$$

To demonstrate, if in Eq. (3.15) we change the variables, $x = t - \tau$, then

$$\int_{-\infty}^{\infty} u(t)\,u(t-\tau)\,dt = \int_{-\infty}^{\infty} u(x+\tau)\,u(x)\,dx$$

An important property of the autocorrelation function is that *under any time shift τ the magnitude of the autocorrelation function does not exceed the signal energy*:

$$|K_u(\tau)| \leqslant K_u(0) = E_u \tag{3.18}$$

This fact is proved by directly using the Cauchy-Buniakovski inequality (about which we have learned in Chap. 1):

$$|(u,\ u_\tau)| \leqslant \|u\| \cdot \|u_\tau\| = E_u \tag{3.19}$$

To sum up, the autocorrelation function $K_u(\tau)$ is a symmetric curve with a central maximum which is always positive. Irrespective of the form of the signal $u(t)$, the autocorrelation function may display both a monotonically decreasing and an oscillating behaviour.

Example 3.3. *The autocorrelation function of a rectangular video pulse.*

Figure 3.4*a* shows a rectangular video pulse of amplitude U and of duration τ_p. It also shows the signal's replica translated backwards in time for τ seconds. The integral of (3.15) is evaluated here in a trivial manner, proceeding from a graphical construction. To demonstrate, the product $u(t)\,u(t-\tau)$ is non-zero only within the time interval over which the superposition of the signals is observed. From Fig. 3.4*a* it is seen that this time interval is equal

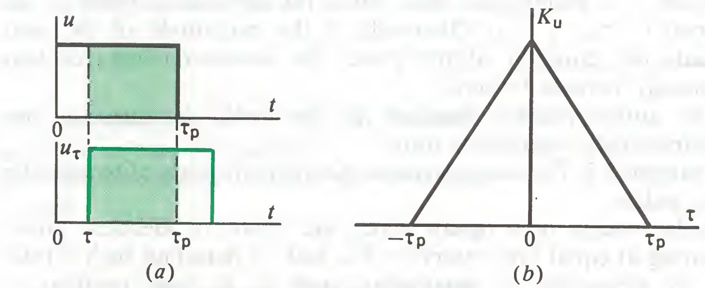


Fig. 3.4 Finding the autocorrelation function of a rectangular video pulse

to $\tau_p - |\tau|$, if the shift does not exceed the duration of the pulse. For the signal in question,

$$K_u(\tau) = \begin{cases} U^2\tau_p\left(1 - \dfrac{|\tau|}{\tau_p}\right), & |\tau| < \tau_p \\ 0, & |\tau| > \tau_p \end{cases} \tag{3.20}$$

Graphically, the function yields a triangle like that shown in Fig. 3.4*b*. The base of the triangle is twice the pulse duration.

Example 3.4. *The autocorrelation function of a rectangular radio pulse.*

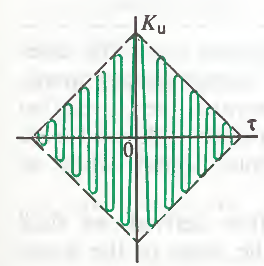
Consider a radio signal of the form

$$u(t) = \begin{cases} 0, & (t < -\tau_p/2) \\ U\cos\omega t, & (-\tau_p/2 < t < \tau_p/2) \\ 0, & (t > \tau_p/2) \end{cases}$$

Since we know that the autocorrelation function has the property of even symmetry, we evaluate the integral in (3.15) by assuming that $0 < \tau < \tau_p$. Then

$$K_u(\tau) = U^2 \int_{-(\tau_p/2)+\tau}^{\tau_p/2} \cos\omega t\,\cos\omega(t-\tau)\,dt$$

$$= (U^2/2)(\tau_p - \tau)\cos\omega\tau + (U^2/2) \int_{-(\tau_p/2)+\tau}^{\tau_p/2} \cos(2\omega t - \tau)\,dt$$



The last trigonometric integral is easy to evaluate, and we obtain

$$K_u(\tau) = (U^2/2)(\tau_p - |\tau|)\left[\cos\omega\tau + \frac{\sin 2\omega(\tau_p - |\tau|)}{2\omega(\tau_p - |\tau|)}\right] \tag{3.21}$$

Naturally, at $\tau = 0$, $K_u(0)$ becomes equal to the energy of the pulse (see Example 1.9). Equation (3.21) defines the autocorrelation

function of a rectangular radio pulse for all shifts τ lying in the interval $(-\tau_p<\tau<\tau_p)$. Obviously, if the magnitude of the shift exceeds the duration of the pulse, the autocorrelation function identically reduces to zero.

The autocorrelation function of the pulse in question has a characteristic oscillatory form.

Example 3.5. *The autocorrelation function of a train of rectangular video pulses.*

Radar widely uses signals which are trains of identical pulses recurring at equal time intervals. The task of detecting such a train and of measuring its parameters, such as its time position, is accomplished by devices which implement the algorithm for the computation of the autocorrelation function of the signal.

As an example, Fig. 3.5*a* shows a train of three identical rectangular video pulses. It also shows its autocorrelation function found by Eq. (3.15) (Fig. 3.5*b*).

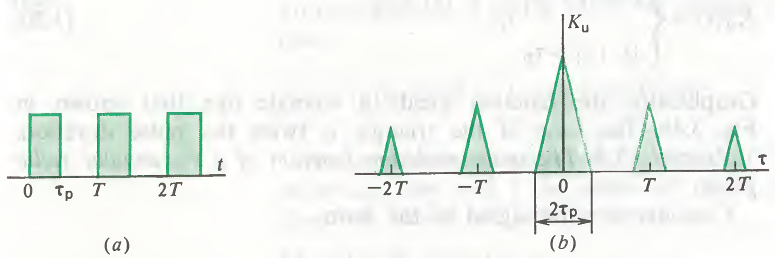


Fig. 3.5 (*a*) A train of three identical video pulses and (*b*) its autocorrelation function

We omit all computational details because the procedure is exactly the same as used in Example 3.3.

It is well seen that the autocorrelation function has a maximum at $\tau=0$. However, when the delay time is a multiple of the pulse repetition period ($\tau=\pm T$, $\pm 2T$ in our case), the plot of the autocorrelation function displays side lobes comparable in height with the main lobe. Therefore, we may speak of a certain imperfection in the correlation structure of the signal.

Work Problem 8

The autocorrelation function of an unbounded signal. In cases involving periodic sequences extended in time without bound, a somewhat different approach is necessary towards the correlation properties of signals. Assume that such a sequence is derived from some signal localized in time, that is, a pulse signal, when its duration τ_p goes to infinity.

To avoid discrepancy in the expressions thus derived, we shall define the new autocorrelation function as the *mean* of the scalar

product of the signal and its time-shifted replica:

$$\tilde{K}_u(\tau)=\lim_{\tau_p\to\infty}\frac{1}{\tau_p}\int_{-\tau_p/2}^{\tau_p/2} u(t)\,u(t-\tau)\,dt \qquad (3.22)$$

With this approach, the autocorrelation function $\tilde{K}_u$ becomes equal to the mean cross power of the two signals.

If we wish to find this autocorrelation function for a cosine waveform unbounded in time

$$u(t)=U\cos\omega t,\quad -\infty<t<\infty$$

we may use Eq. (3.21) derived for a radio pulse of duration τ_p, then pass to the limit with τ_p tending to infinity, and recall the definition in (3.22). As a result, we obtain

$$\tilde{K}_u(\tau)=(U^2/2)\cos\omega\tau \qquad (3.23)$$

This autocorrelation function is itself periodic; its value for $\tau=0$, equal to $U^2/2$, is the mean (effective) power that a given signal will develop across a load resistor of 1 Ω.

Relation between the power spectrum of a signal and its autocorrelation function. From the material presented in this chapter, the student may think that the methods of correlation analysis are only some special techniques which have no direct bearing on the principles of a spectral decomposition. It is easy to show that there is a close relation between the autocorrelation function and the power spectrum of a signal.

To demonstrate, in accord with Eq. (3.15), the autocorrelation function is a scalar product:

$$K_u(\tau)=(u,\ u_\tau)$$

Here u_τ stands for $u(t-\tau)$ which is a time-shifted replica of the original signal.

Referring to the generalized Rayleigh formula, Eq. (3.4), we may write

$$(u,\ u_\tau)=\frac{1}{2\pi}\int_{-\infty}^{\infty} S_u(\omega)\,S_{u_\tau}^*(\omega)\,d\omega$$

The spectrum of a signal shifted in time is

$$S_{u_\tau}=S_u\exp(-j\omega\tau)$$

Therefore,

$$S_{u_\tau}^*=S_u^*\exp(j\omega\tau)$$

This brings us to an important result:

$$K_u(\tau)=\frac{1}{2\pi}\int_{-\infty}^{\infty}|S_u|^2\exp(j\omega\tau)\,d\omega \qquad (3.24)$$

Relation between the autocorrelation function and power spectrum of a signal

As will be recalled, the square of the magnitude of the spectrum is the power spectrum of a signal. Thus, the power spectrum is the Fourier transform of the autocorrelation function of a signal:

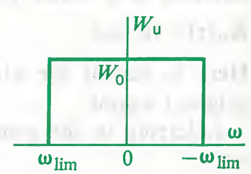
$$K_u(\tau) \leftrightarrow |S_u(\omega)|^2 = W_u(\omega) \tag{3.25}$$

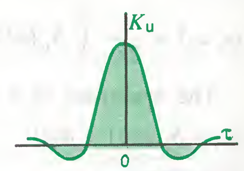
Using Eq. (3.24), we may write the inverse relation

$$|S_u(\omega)|^2 = \int_{-\infty}^{\infty} K_u(\tau)\exp(-j\omega\tau)\,d\tau \tag{3.26}$$

The above expressions are important for two reasons. Firstly, they permit us to evaluate the correlation properties of signals from the distribution of their power over the spectrum. According to the uncertainty principle (see Chap. 2), an increase in the frequency band within which the spectral components of a signal are distributed produces a narrower main lobe of the autocorrelation function and makes the signal more manageable from the view-point of measuring the instant when it occurs. Secondly, Eqs. (3.24) and (3.26) suggest the way in which the power spectrum of a signal can be determined experimentally. Frequently, it is more convenient first to find the autocorrelation function, and then to derive the power spectrum of a signal from its Fourier transform. This procedure is widely used in the analysis of signals with the aid of high-speed computers operating in real time.

Example 3.6. *The autocorrelation function of a signal with a uniform and band-limited power spectrum.*





Let the signal $u(t)$ be characterized by a power spectrum of the form

$$W_u(\omega) = \begin{cases} 0, & \omega < -\omega_{lim} \\ W_0, & -\omega_{lim} < \omega < \omega_{lim} \\ 0, & \omega > \omega_{lim} \end{cases}$$

Using Eq. (3.24), we find the autocorrelation function:

$$K_u(\tau) = (W_0/2\pi)\int_{-\omega_{lim}}^{\omega_{lim}} \exp(j\omega\tau)\,d\omega$$

$$= (W_0/\pi)\int_0^{\omega_{lim}} \cos\omega\tau\,d\omega$$

$$= \frac{W_0\omega_{lim}}{\pi}\,\frac{\sin\omega_{lim}\tau}{\omega_{lim}\tau} \tag{3.27}$$

Thus, the autocorrelation function of the signal in question has a lobed structure.

Correlation time

Frequently, it is convenient to introduce a numerical parameter known as the *correlation time*, τ_c, which is an estimate of the width of the main lobe of the autocorrelation function. It is easy to see that in the case at hand the correlation time should be found from the relation

$$\omega_{lim}\tau_c = \pi$$

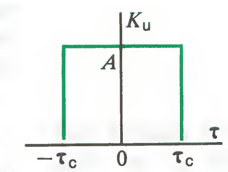
Hence, it follows that

$$\tau_c = \pi/\omega_{lim} = 1/2f_{lim} \tag{3.28}$$

decreases with an increase in the upper frequency limit of the power spectrum of the signal.

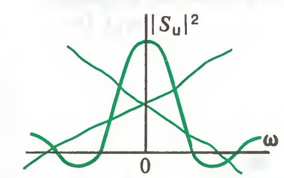
Work Problems 6 and 7

Constraints imposed on the form of the autocorrelation function of a signal. From the relation existing between the autocorrelation function and the power spectrum of a signal we can establish an interesting, but, it seems, not so obvious criterion for the signal to have the desired correlation properties. The point is that the power spectrum $W_u(\omega)$ of any signal must, by definition, be positive (see for example Eq. (3.25)). This condition is not satisfied with just any choice of the autocorrelation function. For example, if we choose



$$K_u(\tau) = \begin{cases} 0, & \tau < -\tau_c \\ A, & -\tau_c < \tau < \tau_c \\ 0, & \tau > \tau_c \end{cases}$$

and take the Fourier transform, then we shall get



$$|S_u|^2 = 2A\int_0^{\tau_c}\cos\omega\tau\,d\tau = (2A/\omega)\sin\omega\tau_c$$

This alternating function cannot represent the power spectrum of any signal.

3.3 The Autocorrelation Function of Discrete Signals

From our discourse on the autocorrelation function of a train of rectangular video pulses, the reader has undoubtedly noted that its plot has a specific lobed structure. For practical uses of the autocorrelation function, such as the detection of this signal or the measurement of its parameters it is immaterial that the lobes are triangular in shape. What is important is their level as compared with the central maximum at $\tau = 0$.

Our immediate task is to modify the autocorrelation function so

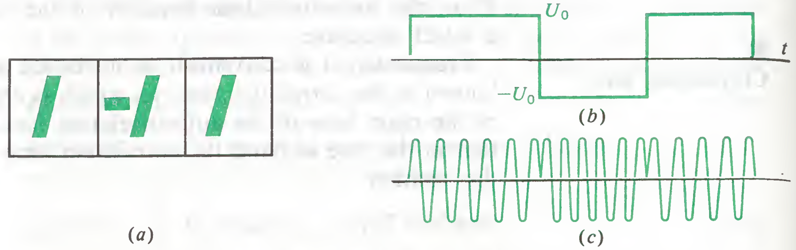


Fig. 3.6 A three-position discrete signal and its realization by variations in amplitude and phase: (*a*) symbolic representation; (*b*) amplitude coding; (*c*) phase coding

that we can extract from it all useful information while divorcing ourselves from minor details. A convenient basis for this endeavour is supplied by the mathematical model of a discrete signal (see Chap. 1).

Principles of discrete signal coding

Representation of composite signals having a discrete structure. A train of identical rectangular video pulses is the simplest representative of the class of composite signals built in accordance with the following principle. The entire interval over which the signal exists is divided into a whole number $M > 1$ of equal intervals called *positions*. Within each position the signal may exist in any one of two physical states.

Figure 3.6 explains several ways of forming a multiposition composite signal. For definiteness (here and elsewhere) we assume that the two possible states of the signal correspond to the numbers $+1$ and -1.

As follows from the figure, a discrete signal may take on different physical shapes. In the first case (see Fig. 3.6*b*) the $+1$ symbol corresponds to the positive value, $+U_0$, of the height of the video pulse transmitted in the respective position, whereas the -1 symbol corresponds to the negative value of the pulse height, $-U_0$. This is the *amplitude coding* of a discrete signal.

In the second case (see Fig. 3.6*c*), the signal is *phase-coded*. In order to transmit the $+1$ symbol in the corresponding position, we generate a truncated harmonic signal with a particular (say, zero) initial phase. In transmitting the -1 symbol, we utilize a truncated sine wave of the same duration and of the same frequency, but with the initial phase shifted through 180°.

Despite the obvious difference between the waveforms of these two signals, we can establish a complete identity between them in terms of their mathematical models. To demonstrate, the model of any such signal is a sequence

$\{u_1, u_2, \ldots, u_{M-1}, u_M\}$

in which each symbol u_j takes on one of two possible values, $+1$ or -1. For convenience, let us agree to stuff the sequence with zeros in the positions where the signal is not defined. Then, the discrete signal $\{1, 1, -1, 1\}$ can be written out in full as

$\ldots\ 0\ 0\ 0\ 1\ 1\ -1\ 1\ 0\ 0\ \ldots$

An important operation in processing discrete signals is the shift of such a signal through a certain number of positions (digits) relative to the original one. In what follows, the first line represents the original signal; the second, third and fourth lines represent the same signal shifted backwards one, two, and three digits:

$\ldots\ 0\ 0\ 0\ 1\ 1\ 1\ 1\ 0\ 0\ 0\ \ldots$
$\ldots\ 0\ 0\ 0\ 0\ 1\ 1\ 1\ 1\ 0\ 0\ 0\ \ldots$
$\ldots\ 0\ 0\ 0\ 0\ 0\ 1\ 1\ 1\ 1\ 0\ 0\ \ldots$
$\ldots\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 1\ 1\ 1\ 0\ \ldots$

The discrete autocorrelation function. Let us generalize Eq. (3.15) so that we could determine the discrete counterpart of the autocorrelation function applicable to the class of signals in question. Obviously, integration should now be replaced with summation, and the variable τ should be replaced with an integer n, positive or negative, which indicates how many positions (digits) the replica is shifted relative to the original signal. Recalling that we have agreed to stuff "empty" positions with zeros, the discrete autocorrelation function may be written as

This formula is applicable to a discrete signal of the most general type

$$\hat{K}_u(n) = \sum_{j=-\infty}^{\infty} u_j u_{j-n} \qquad (3.29)$$

This function of a whole-valued argument n extends to the discrete case the basic ideas of the scalar product of two signals and, quite naturally, possesses many of the already known properties of the scalar product and of the continuous autocorrelation function. For example, it is easy to see that the discrete autocorrelation function shows even symmetry:

$$\hat{K}_u(n) = \hat{K}_u(-n) \qquad (3.30)$$

Under a shift of zero, the discrete autocorrelation function reduces to the energy of a discrete signal:

$$\hat{K}_u(0) = \sum_{j=-\infty}^{\infty} u_j^2 = \hat{E}_u \qquad (3.31)$$

Selected examples. To illustrate the foregoing, let us find the discrete autocorrelation function of a three-digit signal in which all the digits are the same, $u = \{1, 1, 1\}$. Let us write this signal along

with its replicas shifted one, two and three digits:

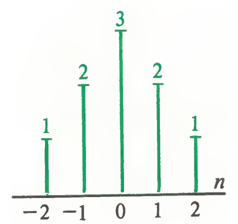
...0 0 0 1 1 1 0 0 0...

...0 0 0 0 1 1 1 0 0...

...0 0 0 0 0 1 1 1 0...

...0 0 0 0 0 0 1 1 1...

As is seen, even with $n = 3$ the original signal and its replica do not overlap any longer. This implies that the products in Eq. (3.29) reduce to zero at $n \geqslant 3$. On taking the sums, we get

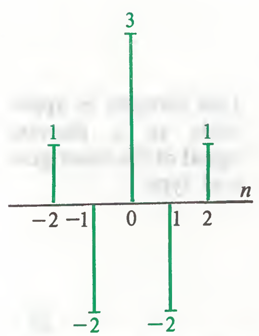


$\hat{K}_u(0) = 1 + 1 + 1 = 3$

$\hat{K}_u(1) = 1 + 1 = 2$

$\hat{K}_u(2) = 1$

The side lobes of the autocorrelation function here subside with increasing number n in about the same manner as we have seen in the case of the autocorrelation function for three continuous video pulses. Obviously, one and the same discrete autocorrelation function can fit a wide variety of sequences derived from any continuous signals, provided the time spacing between the pulses is the same.



Now we consider a discrete signal which differs from the previous one in the sign in the second position:

$u = \{1, -1, 1\}$

Proceeding as before, we find the values of the discrete autocorrelation function:

$\hat{K}_u(0) = 1 + 1 + 1 = 3$

$\hat{K}_u(1) = -1 - 1 = -2$

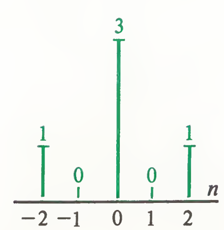
$\hat{K}_u(2) = 1$

It turns out that the first side lobe changes sign, but remains unchanged in magnitude.

Finally, let us consider a three-position discrete signal for which the mathematical model is:

$u = \{1, 1, -1\}$

Its autocorrelation function is:



$\hat{K}_u(0) = 1 + 1 + 1 = 3$

$\hat{K}_u(1) = 1 - 1 = 0$

$\hat{K}_u(2) = 1$

Undoubtedly, of the three discrete signals we have just examined,

it is the third one that is most favourable in terms of correlation properties, because the side lobes of its autocorrelation function have the lowest level.

Barker sequences. In the 50s and 60s, the search for discrete signals whose autocorrelation function would have the best structure was the object of intensive research by specialists in the fields of communication theory and applied mathematics [21]. They have found several classes of signals having very good correlation properties. Of them, the so-called Barker sequences are most known. They possess a unique property: Whatever the number M of positions (digits or pulses), their autocorrelation functions, as found by Eq. (3.29), do not exceed unit for any $n \neq 0$. On the other hand, their energy, that is, $\hat{K}_u(0)$, is numerically equal to M.

It has been found that Barker sequences can be realized only if the number of digits (that is, the length) is $M = 2, 3, 4, 5, 7, 11$, or 13. $M = 2$ represents a trivial case. The Barker sequence with $M = 3$ has just been examined at the end of the previous section. The mathematical models of Barker sequences and the corresponding autocorrelation functions are summarized in Table 3-2.

Table 3-2 Barker Sequences

M	Signal model	Autocorrelation function
3	1, 1, −1	3, 0, −1
4	1, 1, 1, −1	4, 1, 0, −1
	1, 1, −1, 1	4, −1, 0, 1
5	1, 1, 1, −1, 1	5, 0, 1, 0, 1
7	1, 1, 1, −1, −1, 1, −1	7, 0, −1, 0, −1, 0, −1
11	1, 1, 1, −1, −1, −1, 1, −1, −1, 1, −1	11, 0, −1, 0, −1, 0, −1, 0, −1, 0, −1
13	1, 1, 1, 1, 1, −1, −1, 1, 1, −1, 1, −1, 1	13, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1

With $M = 4$, two forms of signals are possible

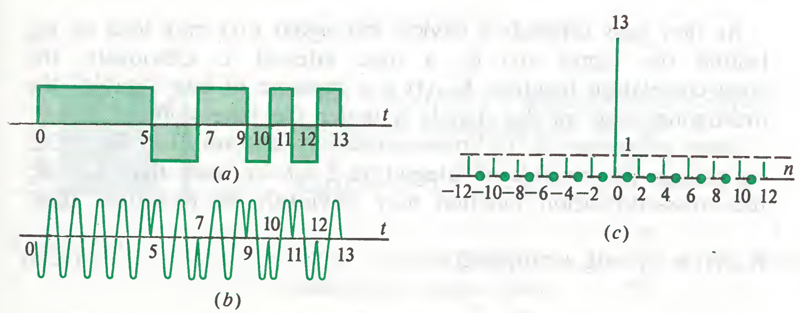


Fig. 3.7 Barker sequences for $M = 13$: (a) amplitude coding; (b) phase coding; (c) autocorrelation function

As an illustration, Fig. 3.7 shows the form of the most commonly used 13-digit Barker sequence and two coding methods along with a plot of its autocorrelation function.

As investigations have shown, Barker sequences with an odd number of digits exceeding 13 are non-existent. It remains unknown, however, whether we can build a Barker sequence with an even M greater than four.

▲ Solve Problem 9

In conclusion, it should be noted that in this chapter we have examined some properties of discrete signals and of their autocorrelation functions by way of an introduction. A systematic exposition of the material will be given in Chap. 15.

3.4 The Cross-Correlation Function of Two Signals

In some theoretical and applied problems of telecommunications it is convenient to introduce a special characteristic of a two-signal system, the *cross-correlation function* between the signals involved. It describes in a unified manner both the difference in shape and the relative position of the signals on the time axis.

Determination of the cross-correlation function. By generalizing Eq. (3.15), let us call the cross-correlation function between two real signals, $u(t)$ and $v(t)$, the scalar product of the form

$$K_{uv}(\tau) = \int_{-\infty}^{\infty} u(t)v(t-\tau)\,dt \tag{3.32}$$

The usefulness of this integral characteristic of signals will be seen from the following example. Let, for instance, the signals $u(t)$ and $v(t)$ be orthogonal in their original state, so that

$$\int_{-\infty}^{\infty} u(t)v(t)\,dt = 0$$

As they pass through a device, the signal $u(t)$ may lead or lag behind the signal $v(t)$ by a time interval τ. Obviously, the cross-correlation function $K_{uv}(\tau)$ is a measure of how "stable" the orthogonal state of the signals is under the time shift.

Some properties of the cross-correlation function. If in Eq. (3.32) we change the variable of integration $\xi = t - \tau$, such that $dt = d\xi$, the cross-correlation function may obviously be re-written thus:

$$K_{uv}(\tau) = \int_{-\infty}^{\infty} u(\xi+\tau)v(\xi)\,d\xi \tag{3.33}$$

The reason why the results found by Eqs. (3.32) and (3.33) are the same is clear: The signals $u(t)$ and $v(t)$ will take up the same relative position, whether we shift the signal $v(t)$ backwards or

advance the signal $u(t)$ by the same amount τ. Therefore,

$$K_{uv}(\tau) = K_{vu}(-\tau) \tag{3.34}$$

In contrast to the autocorrelation function of one signal, the cross-correlation function between two signals does not show even symmetry about the argument τ:

$$K_{uv}(\tau) \neq K_{uv}(-\tau)$$

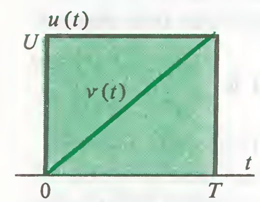
If we consider signals of finite energy, we will see that the cross-correlation function is bounded. This statement follows from the Cauchy-Buniakovski inequality:

$$|K_{uv}(\tau)| = |(u, v_\tau)| \leqslant \| u \| \cdot \| v_\tau \|$$

Hence,

$$|K_{uv}(\tau)| \leqslant \| u \| \cdot \| v \| \tag{3.35}$$

because the shift in time does not affect the norm of the signal. It should be noted that at $\tau = 0$ the cross-correlation function need not reach an absolute maximum.



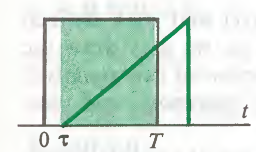
Example 3.7. *Calculate the cross-correlation function $K_{uv}(\tau)$. The signal $u(t)$ is a rectangular video pulse, and the signal $v(t)$ is a triangular video pulse. They have the same amplitude, U, and the same duration, T. Originally (in the absence of delay), the signals exist over a common time interval $[0, T]$.*

For $(0 < t < T)$, the signals can be written:

$$u(t) = U \text{ and } v(t) = Ut/T$$

If $\tau > 0$, that is, if the signal $v(t)$ lags behind the signal $u(t)$, then

$$K_{uv}(\tau) = (U^2/T)\int_0^T (t-\tau)\,dt$$



Case $\tau > 0$

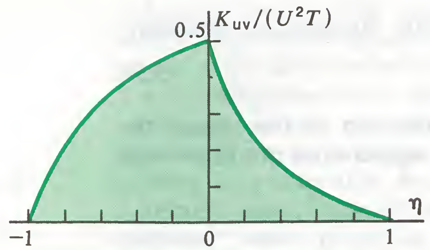
On introducing a dimensionless parameter $\eta = \tau/T$ and carrying out simple manipulations, we get

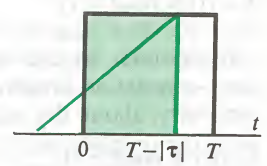
$$K_{uv}(\tau) = U^2 T\left(\frac{1}{2} - \eta + \frac{1}{2}\eta^2\right) \text{ for } \tau > 0 \tag{3.36}$$

If, on the other hand, $\tau < 0$, that is, the triangular pulse leads the rectangular pulse, then

$$K_{uv}(\tau) = (U^2/T)\int_0^{T-|\tau|} (t - |\tau|)\,dt = (U^2T/2)(1-\eta^2) \tag{3.37}$$

The cross-correlation function as found by Eqs. (3.36) and (3.37)





Case $\tau < 0$

Fig. 3.8 A plot of the cross-correlation function between a rectangular and a triangular pulse

is shown diagrammatically in Fig. 3.8. The plot is asymmetrical, because the overlapping area of the two pulses varies at a different rate, depending on the direction of the time shift.

Relation to the cross-power spectrum. Let us express the cross-correlation function between two signals in terms of their spectra. In doing this, we will apply the same reasoning as in the spectral representation of the autocorrelation function. On the basis of the generalized Rayleigh formula

$$K_{uv}(\tau) = (u, v_\tau) = (1/2\pi) \int_{-\infty}^{\infty} S_u(\omega) S^*_{v_\tau}(\omega)\, d\omega$$

Since the spectrum of the time-shifted signal is

$$S_{v_\tau}(\omega) = S_v(\omega) \exp(-j\omega\tau)$$

it follows that

$$K_{uv}(\tau) = (1/2\pi) \int_{-\infty}^{\infty} S_u(\omega) S^*_v(\omega) \exp(j\omega\tau)\, d\omega \tag{3.38}$$

Recalling that

$$W_{uv}(\tau) = S_u(\omega) S^*_v(\omega) \tag{3.39}$$

is the cross-power spectrum of the signals $u(t)$ and $v(t)$, defined over an infinite frequency interval $-\infty < \omega < \infty$, we may draw the following conclusion: *The cross-correlation function of two signals is the Fourier transform of their cross-power spectrum, and vice versa.*

Interestingly, in contrast to the power spectrum of a single signal, the cross-power spectrum contains some information about the phase of the spectral components at different frequencies. Thus, if the signal spectra are

$$S_u(\omega) = |S_u(\omega)| \exp j\psi_u(\omega)$$

$$S_v(\omega) = |S_v(\omega)| \exp j\psi_v(\omega)$$

then, on the basis of Eq. (3.39), the argument of the cross-power spectrum is defined in terms of the difference between the arguments of the spectra of the signals:

$$W_{uv}(\omega) = |S_u(\omega)| \times |S_v(\omega)| \exp\{j[\psi_u(\omega) - \psi_v(\omega)]\}$$

Generalization to the discrete case. Let the signals $u(t)$ and $v(t)$ be specified in discrete form as sets of samples

$$u = \{\ldots, u_{-1}, u_0, u_1, u_2, \ldots\}$$

$$v = \{\ldots, v_{-1}, v_0, v_1, v_2, \ldots\}$$

recurring at equal intervals T. By analogy with the autocorrelation function of a single signal, the cross-correlation function of two discrete signals may be defined by the formula

The cross-correlation function of discrete signals

$$\hat{K}_{uv}(n) = \sum_{j=-\infty}^{\infty} u_j v_{j-n} \tag{3.40}$$

where n is an integer, positive, negative or zero.

We will demonstrate how this function is calculated, using as an example two four-digit Barker sequences

The cross-correlation function of two Barker sequences

$u = \{1, 1, 1, -1\}$

$v = \{1, 1, -1, 1\}$

If $n > 0$, the signal v lags behind the signal u. Proceeding as in the previous section, we compile a table of the signal u and a set of time-shifted replicas of the signal v:

$\ldots 0\ 0\ 0\ 0\ 1\ 1\quad 1\ -1\quad 0\quad 0\ 0\ 0\ 0 \ldots$

$\ldots 0\ 0\ 0\ 0\ 1\ 1\ -1\quad 1\quad 0\quad 0\ 0\ 0\ 0 \ldots$

$\ldots 0\ 0\ 0\ 0\ 0\ 1\quad 1\ -1\quad 1\quad 0\ 0\ 0\ 0 \ldots$

$\ldots 0\ 0\ 0\ 0\ 0\ 0\quad 1\quad 1\ -1\quad 1\ 0\ 0\ 0 \ldots$

$\ldots 0\ 0\ 0\ 0\ 0\ 0\quad 0\quad 1\quad 1\ -1\ 1\ 0\ 0 \ldots$

Using Eq. (3.40), we obtain

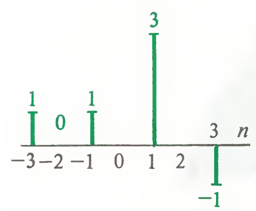
$\hat{K}_{uv}(0) = 0$

$\hat{K}_{uv}(1) = 3$

$\hat{K}_{uv}(2) = 0$

$\hat{K}_{uv}(3) = -1$

A similar table can be compiled with the signal $v(t)$ shifted so



that it leads the signal $u(t)$:

...0 0 0 0 1 1 1 −1 0 0 0...

...0 0 0 0 1 1 −1 1 0 0 0...

...0 0 0 1 1 −1 1 0 0 0 0...

...0 0 1 1 −1 1 0 0 0 0 0...

...0 1 1 −1 1 0 0 0 0 0 0...

Then,

$$\hat{K}_{uv}(-1) = 1$$

$$\hat{K}_{uv}(-2) = 0$$

$$\hat{K}_{uv}(-3) = 1$$

▲ **Work Problem 10**

The plot of the cross-correlation function of these two signals has a sharply asymmetric shape: the cross-correlation function is at a maximum when the signal $v(t)$ is shifted one digit backwards.

Summary

- The scalar product of two signals may be expressed as the scalar product of their spectra (the generalized Rayleigh formula).
- The distribution of the cross power over a frequency range is described by the cross-power spectrum of the two signals involved.
- An approximate orthogonality of signals can be achieved by filtering out appropriate spectral components.
- From the distribution of the signal energy over the entire infinite frequency interval it is seen that its power spectrum is equal to the square of the magnitude of its spectrum.
- The degree of semblance between a signal and its time-shifted replica is described by the autocorrelation function of that signal.
- The power spectrum and the autocorrelation function of a signal are related by a Fourier transform pair.
- The notion of the autocorrelation function can be extended to the case of multiposition (multidigit) discrete signals.
- It is taken that a signal has good correlation properties if the side lobes of its autocorrelation function are substantially smaller than the main lobe.
- The cross-correlation function of two signals is the Fourier transform of their cross-power spectrum.

Review Questions

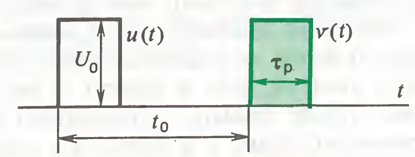
1. What is the physical meaning of the cross-power spectrum of two signals?
2. Name the conditions that should be satisfied by the function describing the cross-power spectrum of two signals so that the signals involved could be orthogonal.

3. Is it possible to realize a situation in which the spectra of two signals overlap, but the signals remain orthogonal?
4. Is the phase of the spectrum of a signal important in determining its power spectrum?
5. Can two nonidentical signals have the same power spectrum?
6. How much of the total energy of a rectangular video pulse does the main (first) lobe of the spectral diagram contain?
7. What are the physical grounds for introducing the autocorrelation function?
8. List the principal properties of the autocorrelation function.
9. What should the power spectrum be for a signal whose autocorrelation function has a narrow main lobe?
10. Define the constraints that can be imposed on the form of the autocorrelation function of a physically realizable signal.
11. What is the basic principle underlying the construction of a multiposition (multidigit) discrete signal?
12. How do we introduce the discrete autocorrelation function of a multiposition signal?
13. Name the main property of Barker sequences. What is the advantage of these sequences over other possible multipulse signals?
14. Can we realize Barker sequences with any number of positions (digits), however large?

Problems

1. Prove that if $u(t)$ and $v(t)$ are real signals, then the imaginary part of the product $S_u(\omega)S_v^*(\omega)$ is an odd function of frequency.

2. Analyse the cross-power spectrum of two identical rectangular video pulses shown in the accompanying diagram



depending on the time shift t_0 between them.

3. Derive an expression describing the power spectrum of the exponential video pulse defined by

$$u(t) = U_0 \exp(-\alpha t)\sigma(t)$$

4. A Gaussian video pulse is defined by

$$u(t) = U_0 \exp(-6 \times 10^{17} t^2)$$

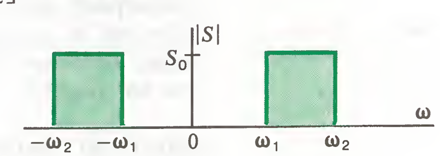
Determine the fraction of the total energy of the signal contained in the frequency range from 0 to 1.5 MHz.

5. Find the effective bandwidth of the exponential signal of Problem 3, assuming it as a frequency band within which 90% of the total signal energy is contained.

6. Prove that the autocorrelation function of the exponential video pulse from Problem 3 is defined by

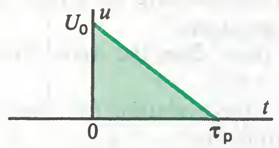
$$K_u(\tau) = (U_0^2/\alpha)\exp(-\alpha|\tau|)$$

7. Find the autocorrelation function of a signal $s(t)$ whose spectrum is real and concentrated in the frequency interval $[\omega_1, \omega_2]$:



8. Calculate the autocorrelation function

of the triangular video pulse shown below:



9. Find the autocorrelation function of the discrete signal $\{1, 1, 1, -1, -1, 1, 1\}$. Compare the result with the autocorrelation function of a seven-digit Barker sequence.

10. Calculate the cross-correlation function of two Barker sequences with $M = 5$ and $M = 7$.

Advanced Problems

11. Analyse experimentally the autocorrelation function of a six-digit discrete signal in which all positions are capable of taking on values $+1$ and -1 with an equal probability. Generate the random numbers by tossing a coin.

Investigate the resulting autocorrelation function by comparing it with that typical of Barker sequences.

Go through the experiment once more for a sufficiently large number of digits (15-20). Suggest the likely approaches to implementing complex signals with good correlation properties in the case of a large number of digits.

12. Find and analyse the cross-correlation function of two exponential signals

$$u(t) = \exp(-\alpha_1 t)\sigma(t)$$

$$v(t) = \exp(-\alpha_2 t)\sigma(t)$$

for $\alpha_1 \neq \alpha_2$.

Chapter 4

Modulated Signals

The baseband signals coming from a message source (a microphone, a TV camera, a telemetry transducer, and the like) cannot, as a rule, be transmitted by a communication channel directly. This is not only because the signals are small in amplitude. The more important factor is that they are relatively low in frequency. For such signals to be effectively transmitted through any medium with the aid of electromagnetic waves, their spectrum must be translated from the low-frequency (baseband) region into one of sufficiently high frequencies. This procedure, known as *modulation* in telecommunications, will be discussed in this chapter.

4.1 Amplitude-Modulated Signals

Before we take up this simplest form of signal modulation in more detail, it is worth while to discuss briefly some matters related to modulation in any form.

The carrier. The rationale of the procedure by which the spectrum of a signal can be translated to higher frequencies is this. To begin with, the transmitter generates an auxiliary r.f. wave called the *carrier*. Its mathematical model

$$u_{car}(t) = f(t;\ a_1,\ a_2,\ \ldots,\ a_m)$$

is such that we can single out a set of parameters $(a^1, a^2, \ldots, a_m)$ which specify the carrier wave.

Let $s(t)$ be the baseband signal to be transmitted by a communication channel. If at least one of the parameters $(a_1, a_2, \ldots, a_m)$ varies in sympathy with the baseband signal, the carrier acquires a new quality—it embodies the information originally contained in the baseband signal $s(t)$.

The physical process by which the parameters of the carrier are varied as desired is what we have called modulation.

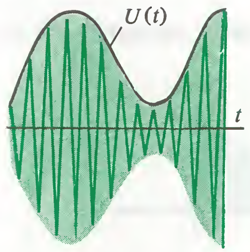
Modulation

Communication systems widely use forms of modulation in which the carrier is a simple harmonic wave

$$u_{car}(t) = U\cos(\omega t + \varphi) \qquad (4.1)$$

In this harmonic wave, we can vary any one of three independent parameters: the amplitude U, the angular frequency ω, or the initial phase φ so that each follows variations in the baseband signal. Depending on which of the three parameters is changed, we can derive three forms of modulation.

The principle of amplitude modulation. When the carrier amplitude $U(t)$ is made to vary in sympathy with the baseband signal while the other two parameters, ω and φ, remain unchanged,



An AM signal and its envelope

we have what is called *amplitude modulation.* An amplitude-modulated (AM) signal can be defined as

$$u_{AM}(t) = U(t)\cos(\omega_0 t + \varphi_0) \tag{4.2}$$

The envelope and the carrier

An AM signal has a specific waveform. Above all, it is symmetrical about the horizontal axis. In accord with Eq. (4.2), an AM signal is the product of the *envelope* $U(t)$ and the *harmonic carrier* $\cos(\omega_0 t + \varphi_0)$. In most cases of practical interest the envelope varies much more slowly than the r.f. carrier.

In amplitude modulation, the envelope $U(t)$ and the modulating baseband signal $s(t)$ are connected by a relation of the form

$$U(t) = U_0[1 + Ms(t)] \tag{4.3}$$

The modulation factor (modulation depth)

Here, U_0 is a constant factor defining the amplitude of the carrier with no modulation applied, and M is called the *modulation factor,* the *depth of modulation* (or *modulation depth*). The value of M defines the degree of amplitude modulation. The meaning of this term can be understood from reference to the waveforms of AM signals in Fig. 4.1.

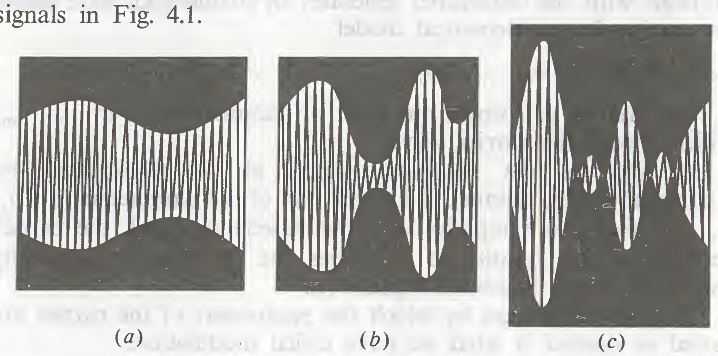


Fig. 4.1 AM signals for different depths of modulation: (*a*) low-level modulation; (*b*) high-level modulation; (*c*) overmodulation

Most radio broadcasting systems operate by amplitude modulation

In the case of a small modulation depth, the relative change in the envelope is small

$$|Ms(t)| \ll 1$$

at all times, irrespective of the waveform of the signal $s(t)$. On the other hand, when the approximate equalities

$$Ms_{max}(t) \approx 1 \quad \text{or} \quad Ms_{min}(t) \approx -1$$

are satisfied at instants when $s(t)$ is at its extremal values, this is an indication of a large modulation depth.

Sometimes, the modulation depth is expressed numerically as a percentage, and is then referred as the *percentage modulation.* This may be the *upward percentage modulation*

$$M_{up} = \frac{U_{max}(t) - U_0}{U_0} \times 100\%$$

or the *downward percentage modulation*

$$M_{down} = \frac{U_0 - U_{min}(t)}{U_0} \times 100\%$$

▲ **Solve Problem 3**

It is not warranted to use AM signals with a small modulation depth in communication channels, because the power output of the transmitter is not fully utilized. In contrast, 100% upward modulation doubles the amplitude of the output wave at the peak values of the modulating signal. Any further increase in amplitude usually leads to undesirable distortion because the output stages of the transmitter are then overloaded.

Overmodulation

An excessive downward modulation is as detrimental. Refer to Fig. 4.1*c* which illustrates what is called *overmodulation* ($M_{down} > 100\%$). Here the envelope ceases to follow the shape of the modulating signal.

Single-tone amplitude modulation. The simplest AM signal is produced when the modulating baseband signal is a harmonic wave at a frequency Ω. Then the resultant (modulated) signal

$$u_{AM}(t) = U_0[1 + M\cos(\Omega t + \Phi_0)]\cos(\omega_0 t + \varphi_0) \tag{4.4}$$

is called a *single-tone AM signal.* It is easy to see that single-tone amplitude modulation has the property of symmetry, that is,

$$M_{up} = M_{down} = M$$

Let us see whether such a signal can be represented as a sum of simple harmonic waves. Using the trigonometric formula for the product of two cosines, we immediately obtain

$$u_{AM}(t) = U_0\cos(\omega_0 t + \varphi_0) + \frac{U_0 M}{2}\cos\left[(\omega_0 + \Omega)t + \varphi_0 + \Phi_0\right] + \frac{U_0 M}{2}\cos[(\omega_0 - \Omega)t + \varphi_0 - \Phi_0] \tag{4.5}$$

Equation (4.5) defines the spectral composition of a single-tone

AM signal. Here

$\omega_0 =$ carrier frequency

$\omega_0 + \Omega =$ upper side frequency

$\omega_0 - \Omega =$ lower side frequency

▲ Work Problem 1

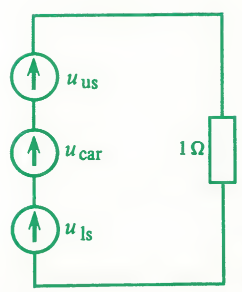
In plotting the spectral diagram of a single-tone AM signal defined in Eq. (4.5), care should be above all taken that the upper and lower side frequencies are equal in magnitude and are arranged symmetrically about the carrier frequency.

The power characteristics of the AM signal. A matter of interest is the distribution of power between the carrier and the side frequencies. The source of a single-tone AM signal is equivalent to a sum of three harmonic sources:

$$u_{car}(t) = U_0 \cos(\omega_0 t + \varphi_0)$$

$$u_{us}(t) = \frac{U_0 M}{2} \cos[(\omega_0 + \Omega)t + \varphi_0 + \Phi_0]$$

$$u_{ls}(t) = \frac{U_0 M}{2} \cos[(\omega_0 - \Omega)t + \varphi_0 - \Phi_0]$$



For definiteness, let us deem that these are ideal voltage sources connected in series and loaded into a 1-ohm resistor. Then the instantaneous power in the AM signal will be numerically equal to the square of the total voltage

$$p(t) = u^2_{AM}(t) = u^2_{car} + u^2_{us} + u^2_{ls} + 2u_{car}u_{us} + 2u_{car}u_{ls} + 2u_{us}u_{ls} \tag{4.6}$$

Characteristically, Eq. (4.6) contains both the self-power of the sources and their cross-powers numerically equal to the pairwise products of instantaneous values.

In order to find the average power of the signal, the term $p(t)$ must be averaged over a sufficiently long time interval T:

$$\bar{p} = \lim_{T\to\infty} \frac{1}{T}\int_0^T p(t)\,dT$$

It can be easily verified that averaging would cause all components of the cross-power to vanish, so the average power of an AM signal is equal to the sum of the average powers of the carrier and the side frequencies:

$$\bar{p} = \bar{p}_{car} + \bar{p}_{us} + \bar{p}_{ls} = U_0^2/2 + U_0^2 M^2/8 + U_0^2 M^2/8 \tag{4.7}$$

Hence, we may write

$$(\bar{p}_{us} + \bar{p}_{ls})/\bar{p}_{car} = M^2/2 \tag{4.8}$$

▲ Work Problem 5

Thus, even in the case of a 100% modulation ($M = 1$), the total power in the two side frequencies is half that of the carrier. Since the information being transmitted is embedded in the side frequencies, we may say that amplitude modulation is rather wasteful of power.

Amplitude modulation by a complex wave. In practice, single-tone AM signals are used but seldom. More frequently, the modulating baseband signal is a complex wave. It is customary to use the following trigonometric sum as a mathematical model for such a signal:

$$s(t) = \sum_{i=1}^{N} \alpha_i \cos(\Omega_i t + \Phi_i) \tag{4.9}$$

Here the frequencies Ω_i form an ordered, ascending sequence $\Omega_1 < < \Omega_2 < \ldots < \Omega_N$, whereas the amplitudes α_i and the initial phases Φ_i are arbitrary.

Substituting (4.9) into (4.3) gives

$$u_{AM}(t) = U_0\left[1 + \sum_{i=1}^{N} M\alpha_i \cos(\Omega_i t + \Phi_i)\right]\cos(\omega_0 t + \varphi_0) \tag{4.10}$$

Let us introduce the set of *partial modulation depths*

● Partial modulation depths

$$M_i = M\alpha_i \tag{4.11}$$

and write an analytical expression for a multitone AM signal in a form which generalizes Eq. (4.4):

$$u_{AM}(t) = U_0\left[1 + \sum_{i=1}^{N} M_i \cos(\Omega_i t + \Phi_i)\right]\cos(\omega_0 t + \varphi_0) \tag{4.12}$$

The spectral expansion is precisely the same as it is for a single-tone AM signal:

$$u_{AM}(t) = U_0\cos(\omega_0 t + \varphi_0) + \sum_{i=1}^{N} \frac{U_0 M_i}{2}\cos[(\omega_0 + \Omega_i)t + \varphi_0 + \Phi_i] + \sum_{i=1}^{N}\frac{U_0 M_i}{2} \times \cos[(\omega_0 - \Omega_i)t + \varphi_0 - \Phi_i] \tag{4.13}$$

The spectral diagram of the modulating signal $s(t)$, plotted in

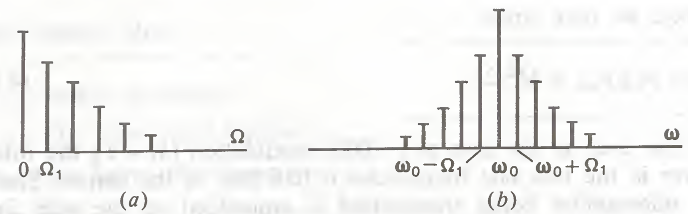


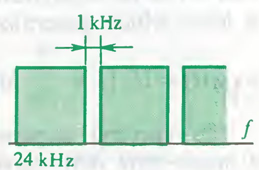
Fig. 4.2 Spectral diagrams: (a) modulating signal; (b) AM signal in the case of multitone modulation

accordance with Eq. (4.9), appears in Fig. 4.2*a*; the spectral diagram of the multitone AM signal corresponding to that baseband signal is shown in Fig. 4.2*b*. As is seen, in addition to the carrier, the spectrum of the multitone AM signal contains a band of upper side frequencies (the *upper sideband*) and a band of lower side frequencies (the *lower sideband*). The spectrum of the upper sideband is a scaled replica of the spectrum of the modulating signal, translated into the high-frequency range by ω_0. The lower sideband, too, has the spectral diagram of the signal $s(t)$, but it is a mirror image of the upper sideband, both sidebands being symmetrically located about the carrier ω_0.

The spectrum of the AM signal

From the foregoing, the following important conclusion may be drawn: *The bandwidth occupied by an AM signal is twice the highest significant frequency in the spectrum of the modulating baseband signal.*

Example 4.1. *Determine the number of radio broadcast channels that can be accommodated in the frequency range from 0.5 to 1.5 MHz (this is approximately the limits of the medium-wave band).*



For radio broadcasts to be reproduced with satisfactory fidelity, it is necessary to reproduce audio frequencies from 100 Hz to 12 kHz. Hence, the bandwidth allotted to each AM channel is 24 kHz. To avoid cross-channel interference, a guard space of, say, 1 kHz, must be provided between adjacent channels. Therefore, the permissible number of channels is

$$N=(1.5-0.5)\times 10^6/(25\times 10^3)=40$$

On-off keyed signals

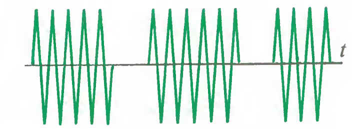
On-off keying. An important class of multitone AM signals are those produced by *keying*. In the simplest case, they are a train of radio pulses separated by intervals within which there is no carrier present. Such signals are typical of radio telegraphy and other systems transmitting discrete information over radio channels.

On-off keyed signals are widely used in pulse radar. For better resolution, use is made of short pulses with a duration of a few fractions of a microsecond.

If $s(t)$ is a function which takes on value 0 or value 1 at each instant of time, then an on-off-keyed signal can be described as

$$u_{\text{keyed}}(t)=U_0\,s(t)\cos(\omega_0 t+\varphi_0) \tag{4.14}$$

As an example, let the function $s(t)$ represent a periodic train of video pulses such as examined in Example 2.1 (see Chap. 2). Assume that their amplitude is $A=1$, then on the basis of Eq. (4.14) we have



Waveform of an on-off keyed signal

$$u_{\text{keyed}}(t)=\frac{U_0}{q}\cos(\omega_0 t+\varphi_0)+\frac{U_0}{q}\sum_{n=1}^{\infty}\frac{\sin(n\pi/q)}{n\pi/q}$$
$$\times\cos[(\omega_0+n\omega_1)t+\varphi_0]+\frac{U_0}{q}\sum_{n=1}^{\infty}\frac{\sin(n\pi/q)}{n\pi/q}$$
$$\times\cos[(\omega_0-n\omega_1)t+\varphi_0] \tag{4.15}$$

From inspection of the above expression, we may conclude that a keyed signal has all the features of a multitone AM signal. The difference, at least theoretical, lies in the fact that its spectrum extends without bound.

Phasor representation of an AM signal. In some cases, it may prove useful to represent an AM signal graphically as the sum of phasors rotating in a complex plane.

For simplicity, we shall limit ourselves to the case of single-tone amplitude modulation. The instantaneous value of the carrier

$$u_{\text{car}}(t)=U_0\cos(\omega_0 t+\varphi_0)$$

is a projection of a phasor, $\dot{U}_{\text{car}}=U_0\exp(j\varphi_0)$, stationary in time, onto the angle axis which rotates about the origin of coordinates at an angular velocity ω_0 in the clockwise direction (Fig. 4.3).

The upper side wave is shown in the diagram as a phasor of length $U_0M/2$, whereas its phase angle at $t=0$ is equal to the sum of the initial phases of the carrier and the modulating signal (see Eq. (4.5)). The lower side wave is represented by a similar phasor which only differs by the sign it takes in the expression for its phase angle. Thus, we should construct on the complex plane the sum of the following three phasors:

$$\dot{U}_{\text{car}}=U_0\exp(j\varphi_0)$$

$$\dot{U}_{\text{us}}=(MU_0/2)\exp[j(\varphi_0+\Phi_0)]$$

$$\dot{U}_{\text{ls}}=(MU_0/2)\exp[j(\varphi_0-\Phi_0)]$$

It is easy to see that the sum will be oriented along the phasor $\dot{U}_{\text{car}}$. The instantaneous value of the AM signal at $t=0$ is equal to

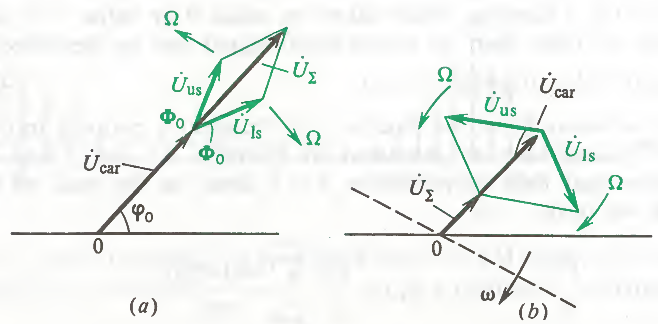


Fig. 4.3 Phasor diagrams of a single-tone AM signal: (a) for $t = 0$; (b) for $t > 0$

the projection of the tip of the resultant phasor on the horizontal axis.

As time goes on, in addition to the rotation of the angle axis we have already spoken about (Fig. 4.3b), two more events will occur: (1) the phasor $\dot{U}_{us}$ will rotate about the point of application at an angular velocity Ω in the counter-clockwise direction, because the total phase angle of the upper side wave $(\omega_0 + \Omega)t + \varphi_0 + \Phi_0$ must increase faster than the phase of the carrier; and (2) the phasor $\dot{U}_{ls}$ will rotate likewise at an angular velocity Ω, but in the opposite direction.

▲ Work Problems 2 and 4

By constructing the phasor $\dot{U}_\Sigma$ and projecting it onto the angle axis, we can determine the instantaneous values $u_{AM}(t)$ at any instant of time.

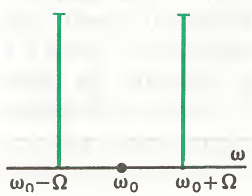
The phasor diagram of a multitone AM signal is identical in meaning, but the graphical construction may be unwieldy.

Double-sideband-suppressed-carrier (DSB-SC) modulation. We have seen that a good proportion of the power in conventional AM signal is concentrated in the carrier. The transmitter power can be better utilized by using AM signals with a suppressed carrier. This is known as *double-sideband-suppressed-carrier (DSB-SC) modulation.* On the basis of Eq. (4.4), a single-tone DSB-SC signal may be represented as follows:

$$u_{SC}(t) = U_0 M \cos(\Omega t + \Phi_0)\cos(\omega_0 t + \varphi_0)$$

$$= (U_0 M/2)\cos[(\omega_0 + \Omega)t + \varphi_0 + \Phi_0]$$

$$+ (U_0 M/2)\cos[(\omega_0 - \Omega)t + \varphi_0 - \Phi_0] \quad (4.16)$$



The spectrum of a single-tone DSB-SC modulated signal

What happens is in effect the multiplication of two signals, the modulating signal and the carrier. From a physical point of view, oscillations such as defined by Eq. (4.16) are produced by *beating* together (heterodyning) two harmonic signals. The beats have the

● Beats

same amplitude $U_0 M/2$ and their frequencies are respectively equal to the upper and lower side frequencies.

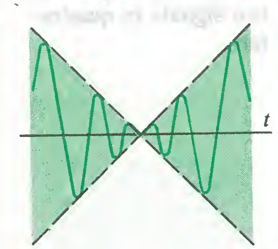
In the case of multitone DSB-SC modulation, the analytical expression for the resultant signal takes the form

$$u_{SC}(t) = U_0 \sum_{i=1}^{N} M_i \cos[(\omega_0 + \Omega_i)t + \varphi_0 + \Phi_i]$$

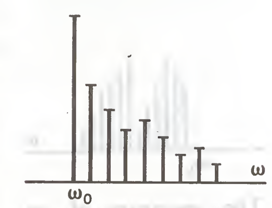
$$+ U_0 \sum_{i=1}^{N} M_i \cos[(\omega_0 - \Omega_i)t + \varphi_0 - \Phi_i] \quad (4.17)$$

As with conventional amplitude modulation, there appear two sidebands, one upper and the other lower, located symmetrically about ω_0.

From an oscillogram of beats, it would appear unclear why there is no carrier in the spectrum of the signal, despite the generation of the carrier frequency component. The explanation is this. As the beat envelope crosses zero, the phase of the r.f. component is abruptly reversed, because the function $(\cos\Omega t + \Phi_0)$ takes opposite signs on the right and left of zero. If such a signal is applied to a high-Q oscillatory system (say, an LC resonant circuit) tuned to frequency ω_0, the output effect will be very small, tending to zero as the Q-factor increases without bound. The oscillations excited in the system by one beat cycle will be cancelled by the next. That is how the spectral expansion of the signal is customarily interpreted from a physical viewpoint. We will go back to this matter in Chap. 9.



For all of its obvious advantages, DSB-SC modulation has not found any appreciable use in radio broadcasting and radio communication. The principal drawback is that in demodulating a DSB-SC signal the carrier must be generated locally at the receiving end and combined with the sidebands. This substantially increases the complexity of the receiver.



The spectrum of an SSB signal

Single-sideband (SSB) modulation. An interesting improvement over the conventional amplitude modulation is the transmission of a signal in which the upper or lower sideband is suppressed. The primary advantage of SSB signals is that they require only half the bandwidth of DSB service. This is of special importance for the frequency-division multiplexing (FDM) of communication channels, such as in operation on short waves where the allotted frequency band is crowded to the limit.

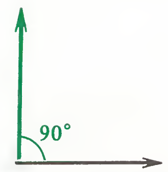
In outward appearance, SSB signals resemble conventional AM signals. For example, a single-tone SSB signal with the lower sideband suppressed may be written as

$$u_{SSB}(t) = U_0\cos(\omega_0 t + \varphi_0) + (U_0 M/2)\cos[(\omega_0 + \Omega)t + \varphi_0 + \Phi_0]$$

On carrying out obvious trigonometric manipulations, we obtain

$$u_{SSB}(t) = U_0 \cos(\omega_0 t + \varphi_0) + (U_0 M/2)\cos(\Omega t + \Phi_0)\cos(\omega_0 t + \varphi_0) - (U_0 M/2)\sin(\Omega t + \Phi_0)\sin(\omega_0 t + \varphi_0)$$
$$= U_0[1 + (M/2)\cos(\Omega t + \Phi_0)]\cos(\omega_0 t + \varphi_0) - (U_0 M/2)\sin(\Omega t + \Phi_0)\sin(\omega_0 t + \varphi_0)$$

The last terms are the products of two functions one of which varies more slowly. Noting that the "faster" terms are in time quadrature with one another, the slowly varying envelope of the SSB signal is found to be



The phasor diagram of two signals in quadrature

$$U(t) = U_0\sqrt{[1 + (M/2)\cos(\Omega t + \Phi_0)]^2 + (M^2/4)\sin^2(\Omega t + \Phi_0)}$$
$$= U_0\sqrt{1 + M\cos(\Omega t + \Phi_0) + M^2/4} \qquad (4.18)$$

A plot of the envelope of an SSB signal, as found by Eq. (4.18) for $M = 1$, appears in Fig. 4.4 which also shows for comparison the envelope of a conventional single-tone AM signal for the same modulation depth.

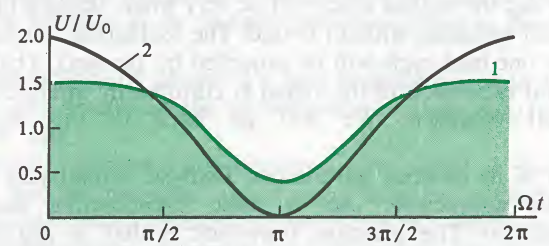
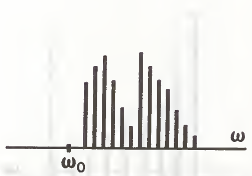


Fig. 4.4 The envelope of a single-tone signal: (*1*) single-tone SSB signal; (*2*) conventional AM signal for $M = 1$



The spectrum of an SSB-SC signal

From comparison of the two curves it is seen that the direct demodulation of an SSB signal from its envelope would introduce appreciable distortion.

A further improvement in the SSB transmission can be achieved by suppressing the carrier partly or completely. In this way, the transmitter power is utilized still better.

4.2 Angle Modulation

Let us consider the case in which the modulating signal $s(t)$ leaves the amplitude U_0 of the harmonic carrier

$$u_{car}(t) = U_0\cos(\omega t + \varphi)$$

unchanged. Instead, the entire argument of the carrier, that is the angle

$$\psi(t) = \omega t + \varphi$$

The total phase

called the *total phase*, varies in sympathy with the modulating signal. Quite aptly, the process is called *angle modulation*.

Principles of angle modulation. Basically, the angle $\psi(t)$ can be varied in one of two ways. In the first of these, the total phase is related to the modulating signal $s(t)$ by an equation of the form

$$\psi(t) = \omega_0 t + ks(t) \qquad (4.19)$$

where k is a proportionality factor, and ω_0 is the unmodulated carrier frequency (existing in the absence of the modulating signal). As follows from Eq. (4.19), the angle $\psi(t)$ varies as the initial phase follows the modulating angle. In this system, the initial phase is proportional to the modulating signal, and the name *phase modulation* (PM) is applied:

Phase modulation

$$u_{PM}(t) = U_0\cos[\omega_0 t + ks(t)] \qquad (4.20)$$

So long as $s(t) = 0$, the PM waveform is a simple harmonic wave. As $s(t)$ builds up in magnitude, the angle $\psi(t)$ increases with time faster than linearly. When the modulating signal decreases, the time rate of change of $\psi(t)$ decreases, too. An example of a PM signal is shown in Fig. 4.5.

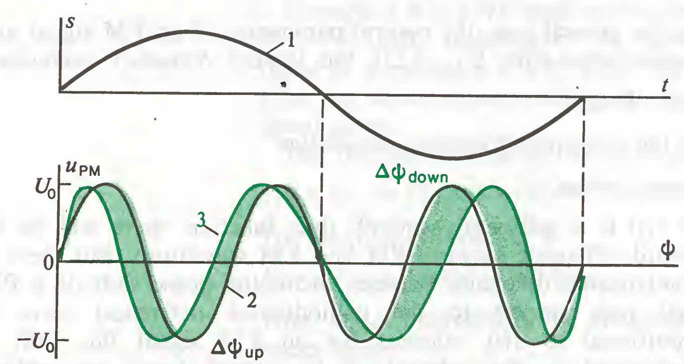


Fig. 4.5 Phase modulation: (*1*) baseband modulating signal; (*2*) unmodulated harmonic wave; (*3*) phase-modulated signal

At the instants of time when $s(t)$ is at its peak value (positive or negative), the absolute value of the phase shift between the PM waveform and the unmodulated carrier is a maximum. This limiting value of the phase shift is called the *phase deviation*, $\Delta\psi$. In the

general case, when the signal $s(t)$ changes sign, it is customary to differentiate between the *upward phase deviation*

Phase deviation

$$\Delta\psi_{up} = ks_{max}$$

and the *downward phase deviation*

$$\Delta\psi_{down} = ks_{min}$$

Instantaneous frequency

If we draw a PM signal on a phasor diagram, we will see that the representative constant-length phasor rotates at a varying angular velocity. The *instantaneous frequency* $\omega(t)$ of an angle-modulated signal is defined as the first time derivative of the total phase angle

$$\omega(t) = \mathrm{d}\psi/\mathrm{d}t \tag{4.21}$$

such that

$$\psi(t) = \int_{-\infty}^{t} \omega(\tau)\,\mathrm{d}\tau + \mathrm{const} \tag{4.22}$$

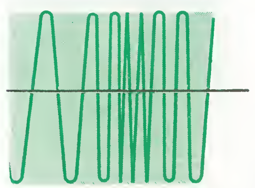
In the second case, the instantaneous frequency $\omega(t)$ of the modulated wave is made proportional to the modulating signal such that

$$\omega(t) = \omega_0 + ks(t) \tag{4.23}$$

Frequency modulation

This is known as *frequency modulation* (FM). Here,

$$u_{FM}(t) = U_0 \cos\left[\omega_0 t + k \int_{-\infty}^{t} s(\tau)\,\mathrm{d}\tau\right] \tag{4.24}$$



The waveform of a typical angle-modulated signal

In the general case, the natural parameters of an FM signal are, in accordance with Eq. (4.23), the *upward frequency modulation*

$$\Delta\omega_{up} = ks_{max}$$

and the *downward frequency modulation*

$$\Delta\omega_{down} = ks_{min}$$

If $s(t)$ is a sufficiently smooth time function, there will be no outward difference between PM and FM waveforms. But there is a fundamental difference between them: the phase shift of a PM signal with respect to the unmodulated (reference) wave is proportional to $s(t)$, whereas for an FM signal this shift is proportional to the integral of the signal being transmitted.

Single-tone angle-modulated signals. Mathematically, it is far more complex to analyse PM and FM signals than AM signals. For this reason, emphasis will be placed on single-tone signals, that is, signals modulated by a single baseband frequency.

For a single-tone FM signal, the instantaneous frequency is given by

$$\omega(t) = \omega_0 + \Delta\omega\cos(\Omega t + \Phi_0)$$

where $\Delta\omega$ is the frequency deviation. On the basis of Eq. (4.22) the total phase of this signal is

$$\psi(t) = \omega_0 t + (\Delta\omega/\Omega)\sin(\Omega t + \Phi_0) + \varphi_0$$

Hence, the quantity

The angle modulation index

$$m = \Delta\omega/\Omega \tag{4.25}$$

called the *angle modulation index* of a single-tone FM signal, represents the deviation of the signal phase, expressed in radians.

Work Problems 6 and 7

To simplify the notation, let us set the time-invariant phase angle φ_0 and Φ_0 equal to zero and write the instantaneous value of the FM signal as

$$u(t) = U_0 \cos(\omega_0 t + m\sin\Omega t) \tag{4.26}$$

The analytical expression for a single-tone PM signal will be the same, but the following point must be borne in mind: PM and FM signals behave differently as the modulating frequency and the amplitude of the modulating signal are varied.

Difference between FM and PM signals

In frequency modulation, the frequency deviation $\Delta\omega$ is proportional to the amplitude of the modulating (baseband) signal, but is independent of the modulating frequency. In phase modulation, the modulation index m is proportional to the amplitude of the modulating signal, irrespective of its frequency. As a consequence, the frequency deviation in FM linearly rises with increasing frequency Ω, in accord with Eq. (4.25).

Example 4.2. *A VHF radio station operating at a carrier frequency of $f_0 = 80$ MHz transmits a PM signal modulated by a baseband signal at a frequency $F = 15$ kHz. The modulation index is $m = 12$. Find the limits between which the instantaneous frequency of the signal varies.*

The mathematical model of the signal has the form

$$u(t) = U_0 \cos[2\pi \times 8 \times 10^7 t + 12 \sin 2\pi \times 1.5 \times 10^4 t]$$

The frequency deviation is

$$\Delta f = mF = 1.8 \times 10^5 = 180 \text{ kHz}$$

Thus, under modulation, the instantaneous frequency of the signal varies from

$$f_{min} = 80 - 0.18 = 79.82 \text{ MHz}$$

to

$$f_{max} = 80 + 0.18 = 80.18 \text{ MHz}$$

Spectral expansion of FM and PM signals at low values of the modulation index. It is relatively simple to represent angle-modulated signals as a sum of harmonic waves, if we limit ourselves to

the case of $m \ll 1$. For this purpose, let us re-arrange Eq. (4.26) in the following way:

$$u(t) = U_0 \cos(\omega_0 t + m \sin \Omega t)$$

$$= U_0 \cos(m \sin \Omega t) \cos \omega_0 t - U_0 \sin(m \sin \Omega t) \sin \omega_0 t \qquad (4.27)$$

Since we assumed a small angle modulation index, we may use the following approximate equalities

$$\cos(m \sin \Omega t) \approx 1$$

$$\sin(m \sin \Omega t) \approx m \sin \Omega t$$

On this basis, we obtain from Eq. (4.27)

$$u(t) \approx U_0 \cos \omega_0 t + (mU_0/2) \cos(\omega_0 + \Omega) t - (mU_0/2) \cos(\omega_0 - \Omega) t \qquad (4.28)$$

The waveforms for which $m \ll 1$ are called narrowband FM or PM signals

Thus, at $m \ll 1$, the spectrum of the angle-modulated signal contains the carrier wave and two side components, upper and lower, at frequencies $\omega_0 + \Omega$ and $\omega_0 - \Omega$. The modulation index m plays here the same role as the modulation depth M in amplitude modulation (cf. Eq. (4.5)). However, we can see a substantial difference between the spectra of an AM signal and an angle-modulated signal. The spectral diagram (see Fig. 4.6*a*) plotted on the basis of Eq. (4.28) is characterized by the fact that the lower side wave has an additional phase shift of 180°.

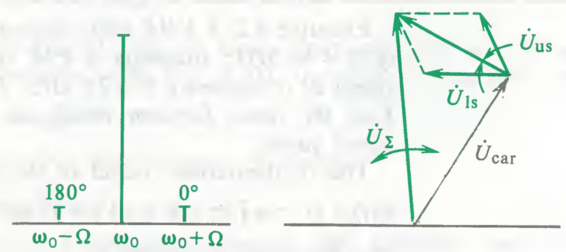


Fig. 4.6 Diagrams of an angle-modulated signal for $m \ll 1$; (*a*) spectral diagram; (*b*) phasor diagram

As a consequence, the sum of the phasors representing the two side waves (see Fig. 4.6*b*) is always at right angles to the carrier phasor $\dot{U}_{car}$. With time, the phasor $\dot{U}_\Sigma$ swings about a central position. The small variations in the length of the phasor are related to the approximate nature of our analysis; at very low values of m they may be justifiably neglected.

A more rigorous spectral analysis of angle-modulated signals. We may try to refine the result obtained above, by using two terms

from the series expansion of harmonic functions of a small argument. Then Eq. (4.27) takes the following form

$$u(t) \approx U_0 \cos \omega_0 t (1 - m^2 \sin^2 \Omega t/2)$$

$$- U_0 \sin \omega_0 t (m \sin \Omega t - m^3 \sin^3 \Omega t/6)$$

Simple trigonometric manipulations yield the following result:

$$u(t) = U_0 (1 - m^2/4) \cos \omega_0 t + U_0 m (1 - m^2/8) [\cos(\omega_0 + \Omega) t$$

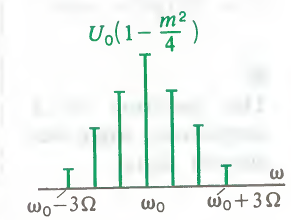
$$- \cos(\omega_0 - \Omega) t] + U_0 (m^2/8) [\cos(\omega_0 + 2\Omega) t$$

$$+ \cos(\omega_0 - 2\Omega) t] + (U_0 m^3/48) [\cos(\omega_0 + 3\Omega) t$$

$$- \cos(\omega_0 - 3\Omega) t] \qquad (4.29)$$

Solve Problem 11

Equation (4.29) reflects a remarkable fact: it states that in addition to the already known components the spectrum of a single-tone angle modulated signal contains the upper and lower side waves at the harmonic frequencies of the modulating wave. Therefore, the spectrum of such a signal is more complex in structure than that of the respective AM signal. Also, the new spectral components lead to a redistribution of energy over the spectrum. As is seen from Eq. (4.29), an increase in m is accompanied by a rise in the amplitude of the side components, whereas the amplitude of the carrier wave decreases by a factor of $(1 - m^2/4)$.



A refined spectral diagram

The spectrum of an angle-modulated signal at an arbitrary modulation index. For the simplest case of single-tone angle modulation we can derive a general expression applicable to any value of the modulation index m.

The division of mathematics dealing with special functions proves an important formula related to the Fourier series expansion – the exponential function with a special form of imaginary exponent:

$$\exp(-ja \sin x) = \cos(a \sin x) - j \sin(a \sin x)$$

$$= J_0(a) + 2 \sum_{k=1}^{\infty} J_{2k}(a) \cos 2kx$$

$$- j2 \sum_{k=1}^{\infty} J_{2k-1}(a) \sin(2k-1) x \qquad (4.30)$$

where $J_k(a)$ is the Bessel function of a, of order k. From a comparison of (4.30) and (4.27), we may re-write the last equation thus:

$$u(t) = U_0 \cos \omega_0 t [J_0(m) + 2 \sum_{k=1}^{\infty} J_{2k}(m) \cos 2k\Omega t]$$

$$- U_0 \sin \omega_0 t [2 \sum_{k=1}^{\infty} J_{2k-1}(m) \sin(2k-1) \Omega t] \qquad (4.31)$$

Since

$$\cos\omega_0 t\cos 2k\Omega t=\frac{1}{2}\cos(\omega_0-2k\Omega)t+\frac{1}{2}\cos(\omega_0+2k\Omega)t$$

and

$$\sin\omega_0 t\sin(2k-1)\Omega t=\frac{1}{2}\cos[\omega_0-(2k-1)\Omega]t$$
$$-\frac{1}{2}\cos[\omega_0+(2k-1)\Omega]t$$

we may re-write Eq. (4.31) in a more compact form

$$u(t)=U_0\sum_{k=-\infty}^{\infty}J_k(m)\cos(\omega_0+k\Omega)t \tag{4.32}$$

The spectrum of a single-tone angle-modulated signal

To sum up, the spectrum of a single-tone angle-modulated signal contains, in the general case, an infinite number of components whose frequencies are $\omega_0 \pm k\Omega$; and their amplitudes are proportional to the values of $J_k(m)$.

In the theory of Bessel functions it is proved that functions with positive and negative indices are interrelated:

$J_{-k}(m)=(-1)^k J_k(m)$

Therefore, the initial phases of the sidebands $\omega_0+k\Omega$ and $\omega_0-k\Omega$ are the same if k is even, and differ by 180° if k is odd.

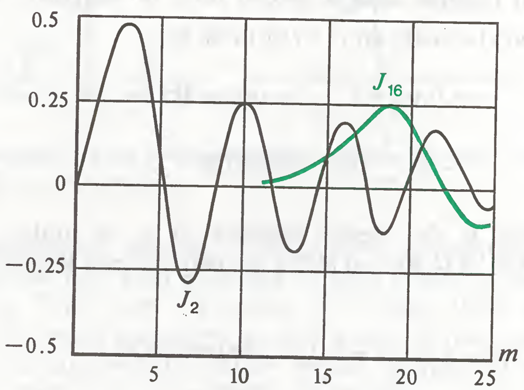


Fig. 4.7 Plots of Bessel functions $J_2(m)$ and $J_{16}(m)$

For a detailed analysis and the construction of spectral diagrams, it is important to know the behaviour of functions $J_k(m)$ according to the value of k for various values of m. Plots of two Bessel functions substantially differing in the value of k are given in Fig. 4.7.

The following trend is apparent: The greater the order of a Bessel function, the greater the range of values of the argument for which the function is very small. More accurately this fact is depicted in Table 4-1.

Table 4-1 Values of Bessel functions $J_k(m)$

k \ m	1	2	3	4	5
0	0.765	0.224	−0.260	−0.397	−0.178
1	0.440	0.577	0.339	−0.066	−0.328
2	0.115	0.353	0.486	0.364	0.047
3	0.020	0.129	0.309	0.430	0.365
4	0.002	0.034	0.132	0.281	0.391
5	2×10^{-4}	0.007	0.043	0.132	0.261
6	2×10^{-5}	0.001	0.011	0.049	0.131
7	1×10^{-6}	2×10^{-4}	0.003	0.015	0.053

In the shaded area the Bessel functions become negligibly small

Using Table 4-1 together with Eq. (4.32), we can readily construct typical spectral diagrams for single-tone angle-modulated signals at not very large values of the modulation index m (Fig. 4.8).

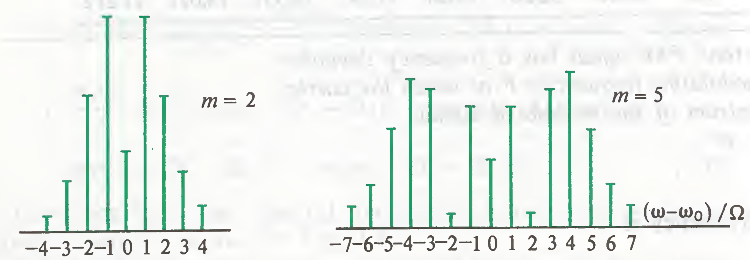


Fig. 4.8 Spectral diagrams of angle-modulated signals for two values of m. (The amplitudes are plotted on a relative scale.)

Interestingly, the increase in the modulation index is accompanied by a broadening in the bandwidth occupied by the signal. Ordinarily, it is assumed legitimate to neglect all spectral components for which $k > m+1$. Hence, the practical bandwidth for an angle-modulated signal may be estimated as

The practical bandwidth of FM and PM signals

$$\mathrm{BW}_{\mathrm{pract}} = 2(m+1)\Omega \tag{4.33}$$

As a rule, for real FM and PM signals $m \gg 1$. Therefore,

$$\mathrm{BW}_{\mathrm{pract}} \approx 2m\Omega = 2\Delta\omega \tag{4.34}$$

Work Problem 12

Or, in words, *an angle-modulated signal occupies a bandwidth approximately equal to twice the frequency deviation.*

As will be recalled, the transmission of AM signals requires a bandwidth equal to 2Ω, or by a factor of m smaller. The broadband nature of FM and PM signals limits their applicability to the metric and shorter wavelengths (VHF and UHF bands). On the other hand, this very property – the broadband nature – makes angle-modulated signals more immune to interference and noise as compared with AM signals.

Angle-modulated signals are often used for high-fidelity broadcasts in the VHF and UHF bands

It has already been noted that the increase in the angle modulation index entails a re-distribution of power within the signal spectrum. Notably, if the value of m is chosen such that

$$J_0(m) = 0 \tag{4.35}$$

the carrier frequency ω_0 will be absent in the spectrum. The values of m, which are the roots of Eq. (4.35), form an infinitely increasing sequence of numbers m_ν ($\nu = 1, 2, \ldots$ is the No. of the root). Several roots of the equation are given in Table 4-2.

Table 4-2 Roots of the equation $J_0(m) = 0$

ν	1	2	3	4	5	6	7
m_ν	2.405	5.520	8.654	11.792	14.931	18.071	21.212

Example 4.3. *A single-tone FM signal has a frequency deviation $\Delta f = 240$ kHz. Find the modulating frequencies F at which the carrier will be absent in the spectrum of the modulated signal.*

The modulation index is

$$m = \Delta\omega/\Omega = \Delta f/F$$

Hence, the modulating frequency is

$$F = \Delta f/m$$

Referring to Table 4-2, we find the sequence of frequencies that

meet the stated condition:

$$F_1 = 240 \div 2.405 = 99.792 \text{ kHz}$$

$$F_2 = 240 \div 5.520 = 43.478 \text{ kHz}$$

$$F_3 = 240 \div 8.654 = 27.732 \text{ kHz}$$

. . . .

Angle modulation in the case of a nonharmonic modulating wave. An interesting behaviour of angle-modulated waves is brought out if we analyse the case where the modulating wave is not harmonic. For simplicity, let us consider a signal modulated by two baseband frequencies:

$$\begin{aligned} u(t) &= U_0 \cos(\omega_0 t + m_1 \sin\Omega_1 t + m_2 \sin\Omega_2 t) \\ &= U_0 \cos\omega_0 t \cos(m_1 \sin\Omega_1 t + m_2 \sin\Omega_2 t) \\ &\quad - U_0 \sin\omega_0 t \sin(m_1 \sin\Omega_1 t + m_2 \sin\Omega_2 t) \end{aligned} \tag{4.36}$$

Assume that the partial modulation indices m_1 and m_2 are so small that we may use approximate expressions for the cosine and the sine:

$$\cos x \approx 1 - x^2/2, \quad \sin x \approx x$$

By carrying out somewhat tedious, but elementary trigonometric manipulations, we may write the original signal as a sum:

$$\begin{aligned} u(t) &= U_0\left(1 - \frac{m_1^2 + m_2^2}{4}\right)\cos\omega_0 t + \frac{m_1 U_0}{2}[\cos(\omega_0 + \Omega_1)t \\ &\quad - \cos(\omega_0 - \Omega_1)t] + \frac{m_2 U_0}{2}[\cos(\omega_0 + \Omega_2)t \\ &\quad - \cos(\omega_0 - \Omega_2)t] + \frac{m_1^2 U_0}{8}[\cos(\omega_0 + 2\Omega_1)t \\ &\quad + \cos(\omega_0 - 2\Omega_1)t] + \frac{m_2^2 U_0}{8}[\cos(\omega_0 + 2\Omega_2)t + \cos(\omega_0 - 2\Omega_2)t] \\ &\quad + \frac{m_1 m_2}{2} U_0 [\cos(\omega_0 + \Omega_1 - \Omega_2)t + \cos(\omega_0 - \Omega_1 + \Omega_2)t \\ &\quad - \cos(\omega_0 + \Omega_1 + \Omega_2)t - \cos(\omega_0 - \Omega_1 - \Omega_2)t] \end{aligned} \tag{4.37}$$

In sketch form, the spectral diagram of a two-tone angle-modulated signal is shown in Fig. 4.9.

A fact of fundamental importance should be stressed: In addition to the frequencies $\omega_0 \pm \Omega_1$, $\omega_0 \pm \Omega_2$, $\omega_0 \pm 2\Omega_1$ and $\omega_0 \pm 2\Omega_2$, the

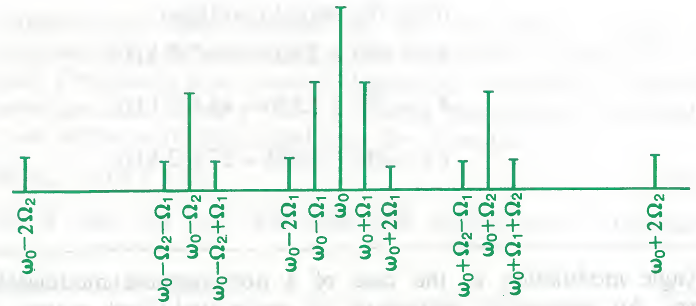


Fig. 4.9 Spectral diagram of a double-tone angle-modulated signal for low values of the partial modulation indices m_1 and m_2

spectrum of the above signal contains *intermodulation products* $\omega_0 \pm \Omega_1 \pm \Omega_2$, taking all the four signs. Their amplitudes depend on the product of the two partial modulation indices.

Work Problem 18

If we abandon the assumption of a small modulation index and assume that the modulation is effected by a group of baseband frequencies $\Omega_1, \Omega_2, \ldots, \Omega_N$, then the spectrum of a phase- or frequency-modulated signal will contain all likely frequencies $\omega_0 \pm \pm n_1\Omega_1 \pm \ldots \pm n_N\Omega_N$, where $n_1, \ldots, n_N$ are all likely integers, including zero. Thus, with all other conditions being equal, the spectrum of PM and FM signals is far richer in components than the spectrum of AM signals.

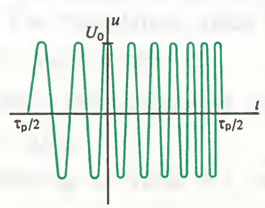
Do not confuse "nonlinear modulation" and "nonlinear network"

In order to stress the interaction between the various components of the modulating signal, angle modulation is sometimes called *nonlinear* in contrast to amplitude modulation.

4.3 Pulsed FM Signals

Now let us examine the spectral and correlation properties of a special class of modulated signals which have recently come to be widely used in radar. They differ from simple r.f. pulses in that their carrier varies (is modulated) in frequency. Most frequently, use is made of a type of pulse frequency modulation in which the instantaneous frequency varies linearly with time.

The principle of linear frequency modulation (LFM). Consider an r.f. pulse having a rectangular envelope, and assume that the r.f. carrier of the pulse has a frequency linearly rising with time. To make the mathematical model of this signal more specific, let its duration be τ_p, such that the point $t=0$ occurs in the middle of the pulse, and the instantaneous frequency varies with time as



$$\omega(t) = \omega_0 + \mu t \tag{4.38}$$

where ω_0 is the carrier frequency and μ is a parameter of dimension

s^{-2}, which defines the time rate of change of the frequency.

It is easy to see that over the time interval equal to the pulse duration the frequency deviation will be

$$\Delta\omega = \mu\tau_p \tag{4.39}$$

The total phase of the signal is

$$\psi(t) = \omega_0 t + \mu t^2/2 \tag{4.40}$$

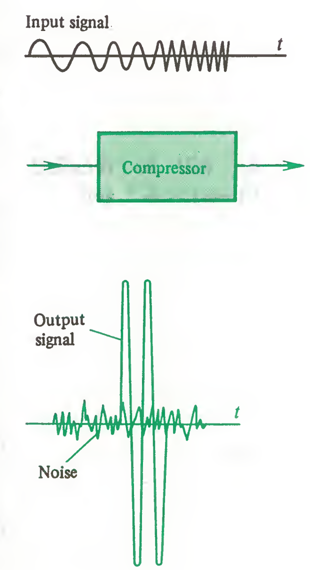
To this we should have added the constant phase shift φ_0, but its presence is immaterial.

Thus, let us use the term *linear-FM radio pulse* (or shortly, *LFM pulse*) to call a signal representable by the following mathematical model:

$$u(t) = \begin{cases} 0, & t < -\tau_p/2 \\ U_0 \cos(\omega_0 t + \mu t^2/2), & -\tau_p/2 < t < \tau_p/2 \\ 0, & t > \tau_p/2 \end{cases} \tag{4.41}$$

The remarkable property of LFM signals that makes them of practical value consists in the following. Suppose that we have at our disposal a physical device which delays the signals applied to its input. If the delay time t_d is arranged to decrease with increasing signal frequency, then, given certain conditions, the long-duration LFM signals applied to the device will be "compressed" in time. More specifically, the output signal from the delay device, or the compressor, will contain both the low-frequency components associated with the start of the applied pulse, and the higher-frequency components associated with its finish at practically the same time owing to the time delay.

Such devices are called dispersive delay lines



It is easy to realize that if the power loss within the compressor is small, the amplitude of the output signal will be boosted appreciably and can markedly exceed the noise level. This enhances the reliability of radar signal detection in noise.

The mechanism of LFM signal compression we have just described gives only an idea about the principle by which such systems operate. A more detailed analysis from which the effect can be judged quantitatively will be given in Chap. 16 in connection with the optimal detection of signals in noise.

The spectrum of a rectangular LFM pulse. In the previous section concerned with the spectral characteristics of an FM signal modulated by two baseband frequencies, we have seen that its spectrum has a complex structure due to intermodulation, or the cross-interaction of the individual components. All this fully applies to the spectrum of an LFM pulse. In the discussion that follows we shall mainly adhere to the notation used in [21].

On the basis of (4.41), the amplitude spectral density, or spectrum

of a single LFM pulse may be written as

$$U(\omega) = U_0 \int_{-\tau_p/2}^{\tau_p/2} \cos(\omega_0 t + \mu t^2/2)\exp(-j\omega t)\,dt$$

$$= (U_0/2) \int_{-\tau_p/2}^{\tau_p/2} \exp\{j[(\omega_0 - \omega)t + \mu t^2/2]\}\,dt$$

$$+ (U_0/2) \int_{-\tau_p/2}^{\tau_p/2} \exp\{-j[(\omega_0 + \omega)t + \mu t^2/2]\}\,dt \tag{4.42}$$

In analysing the above relation, we note that the first integral describes that part of the spectrum which has a sharp maximum in the region of positive frequencies close to ω_0. Accordingly, the second integral corresponds to that part of the spectrum which is mainly concentrated at $\omega < 0$. In practice, we are solely interested in the case where the overlapping of the spectra concentrated at positive and negative frequencies is negligibly small. This is because the total frequency deviation over the pulse duration is a very small fraction of the carrier frequency:

$$\Delta\omega = \mu\tau_p \ll \omega_0$$

Therefore, in Eq. (4.42) only the first integral needs to be evaluated as it defines the spectrum at $\omega > 0$. Then the negative-frequency spectrum may be derived on the basis of the known properties of the Fourier transform for real signals (see Chap. 2).

As will be recalled, $U(-\omega) = U^*(\omega)$

In view of the foregoing and completing the argument of the exponential function in (4.42) to the square, we obtain

$$U(\omega) = (U_0/2)\exp\left[-j\frac{(\omega-\omega_0)^2}{2\mu}\right]\int_{-\tau_p/2}^{\tau_p/2}\exp\left[j(\mu/2)\left(t - \frac{\omega-\omega_0}{\mu}\right)^2\right]dt \tag{4.43}$$

It is convenient to change from the variable t to a new argument, x, in accordance with the formula

$$\sqrt{\mu}\left(t - \frac{\omega-\omega_0}{\mu}\right) = \sqrt{\pi}\,x$$

On calculating, we find

$$U(\omega) = (U_0/2)\sqrt{\pi/\mu}\exp\left[-j\frac{(\omega-\omega_0)^2}{2\mu}\right]\int_{-X_1}^{X_2}\exp(j\pi x^2/2)\,dx \tag{4.44}$$

where the integration limits are

$$X_1 = \frac{\mu\tau_p/2 + (\omega - \omega_0)}{\sqrt{\pi\mu}}$$

and

$$X_2 = \frac{\mu\tau_p/2 - (\omega - \omega_0)}{\sqrt{\pi\mu}} \tag{4.45}$$

The integral in Eq. (4.44) reduces to a combination of well explored special functions called the *Fresnel integrals*:

Fresnel integrals are widely used in physics to solve problems of wave diffraction

$$C(x) = \int_0^x \cos(\pi\xi^2/2)\,d\xi$$

$$S(x) = \int_0^x \sin(\pi\xi^2/2)\,d\xi$$

For $x \gg 1$,

$$C(x) \approx \frac{1}{2} + \frac{\sin(\pi x^2/2)}{\pi x}$$

$$S(x) \approx \frac{1}{2} - \frac{\cos(\pi x^2/2)}{\pi x}$$

As a result, we obtain the final formula for the spectrum of the LFM signal in question:

$$U(\omega) = (U_0/2)\sqrt{\pi/\mu}\exp\left[-j\frac{(\omega-\omega_0)^2}{2\mu}\right] \times \{C(X_1) + C(X_2) + j[S(X_1) + S(X_2)]\} \tag{4.46}$$

If we write the spectrum in exponential form

$$U(\omega) = |U(\omega)|\exp[j\Phi(\omega)]$$

it will be noted that its magnitude

$$|U(\omega)| = (U_0/2)\sqrt{\pi/\mu}\sqrt{[C(X_1) + C(X_2)]^2 + [S(X_1) + S(X_2)]^2} \tag{4.47}$$

whereas the phase spectrum consists of a quadratic term

$$\Phi_1(\omega) = -(\omega - \omega_0)^2/2\mu \tag{4.48}$$

and the so-called residual phase term

$$\Phi_2(\omega) = \arctan\frac{S(X_1) + S(X_2)}{C(X_1) + C(X_2)} \tag{4.49}$$

LFM signals with a large frequency deviation-pulse duration product. Numerical analysis of the above equations indicates that the manner in which the magnitude and phase of the spectrum of a rectangular LFM pulse depend on frequency is fully decided by a dimensionless number

The frequency deviation-pulse duration product

$$B = \Delta\omega\tau_p = \mu\tau_p^2 \tag{4.50}$$

called the *frequency deviation-pulse duration product* of an LFM signal.

In cases of practical interest, $B \gg 1$. The spectrum of such LFM signals has a number of specific properties. Firstly, the magnitude of the spectrum is practically constant over the bandwidth $\Delta\omega$ with centre at ω_0. The corresponding plots based on Eqs. (4.47) and (4.49) appear in Fig. 4.10.

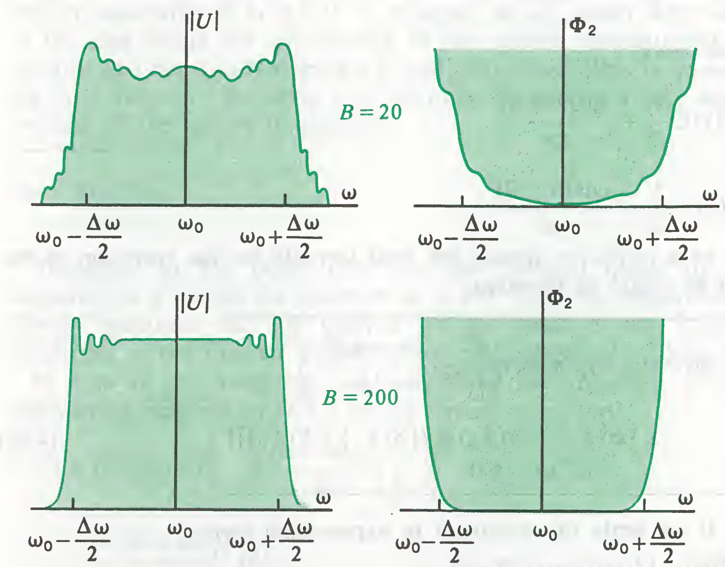


Fig. 4.10 Frequency dependences of the magnitude and the residual phase term of the spectrum of a rectangular LFM signal for several values of the frequency deviation-pulse duration product

Secondly, the oscillations of the magnitude gradually decay with increasing frequency deviation-pulse duration product. From an analysis of (4.47), it can be verified that at the centre frequency of the spectrum

Work Problem 14

$$|U(\omega_0)| = U_0\sqrt{\pi/2\mu} \tag{4.51}$$

Thus, an LFM signal with a large frequency deviation-pulse duration product can approximately be represented by the magnitude of the spectrum

$$|U(\omega)| = \begin{cases} 0, & 0 < \omega < \omega_0 - \Delta\omega/2 \\ U_0\sqrt{\pi/2\mu}, & \omega_0 - \Delta\omega/2 < \omega < \omega_0 + \Delta\omega/2 \\ 0, & \omega > \omega_0 + \Delta\omega/2 \end{cases} \tag{4.52}$$

The power spectral density of such a signal

$$W_u = \pi U_0^2/2\mu \tag{4.53}$$

is likewise constant within the frequency interval $(\omega_0 - \Delta\omega/2,\ \omega_0 + \Delta\omega/2)$, and vanishes elsewhere.

Example 4.4. *A rectangular LFM pulse has an amplitude $U_0 = 20$ V, a carrier frequency $f_0 = 10$ GHz, and a duration $\tau_p = 2$ µs. The frequency deviation over the pulse duration is $\Delta f = 0.1$ GHz. Find the basic parameters of the signal spectrum.*

To begin with, we find the frequency deviation-pulse duration product of the signal:

$$B = 6.28 \times 10^8 \times 2 \times 10^{-6} = 12.56 \times 10^3$$

The rate of frequency rise is

$$\mu = 2\pi\Delta f/\tau_p = 6.28 \times 10^8 \div (2 \times 10^{-6}) = 3.14 \times 10^{14}\ \mathrm{s}^{-2}$$

By Eq. (4.53), the power spectral density is

$$W_u = 0.5 \times 10^{-12}$$

Since the signal has a large frequency deviation-pulse duration product, its spectrum is practically enclosed within the bandwidth from

$$f_0 - \Delta f/2 = 9.95\ \mathrm{GHz}$$

to

$$f_0 + \Delta f/2 = 10.05\ \mathrm{GHz}$$

The autocorrelation function of an LFM signal. In order to find this characteristic so important in signal detection, it is advisable to use the results obtained in Chap. 3 where it is shown that the autocorrelation function and the power spectrum of a signal are uniquely related by a Fourier transform pair.

Let us assume that the frequency deviation-pulse duration product, B, of an LFM signal is so large that the power spectrum of the signal is uniform and concentrated around the carrier frequency within a frequency band $\Delta\omega$ wide. Then the autocorrelation

function of the LFM signal [see Eq. (4.53)] may be defined as

$$K_{\mathrm{LFM}}(\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty} W_u(\omega)\exp(j\omega\tau)\,d\omega$$

$$= (U_0^2/2\mu)\int_{\omega_0-\mu\tau_p/2}^{\omega_0+\mu\tau_p/2} \cos\omega\tau\,d\omega$$

$$= \frac{U_0^2\tau_p}{2}\,\frac{\sin(\mu\tau_p\tau/2)}{\mu\tau_p\tau/2}\cos\omega_0\tau \qquad (4.54)$$

A plot of the normalized autocorrelation function

$r_{\mathrm{LFM}}(\tau) = K_{\mathrm{LFM}}(\omega)/K_{\mathrm{LFM}}(0)$

appears in Fig. 4.11 which also shows the lobed envelope of the function.

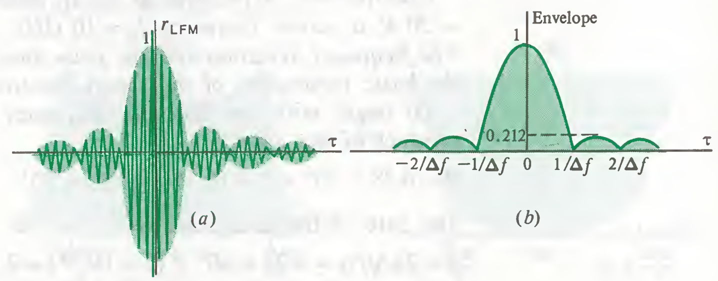


Fig. 4.11 Correlation properties of an LFM-pulse: (*a*) normalized autocorrelation function; (*b*) its envelope

Equation (4.54) defines an important property of LFM signals: The width of the main lobe in the envelope of the autocorrelation function is inversely proportional to the frequency deviation over the pulse duration. This is so because the envelope first reduces to zero when the time shift between the signal and its replica is equal to $\tau = 2\pi/\mu\tau_p = 1/\Delta f$. Radar signals have a substantial frequency deviation, so the main lobe of the autocorrelation function is very narrow. For instance, taking the signal defined in Example 4.4, the time shift at which the envelope of the autocorrelation function reduces to zero is a mere 0.01 μs, which is 0.5% of the pulse duration.

From the view-point of correlation properties, however, LFM signals suffer from a limitation: the height of the first two symmetrical side lobes of the autocorrelation function is fairly large, being 0.212 of that of the main lobe. Given unfavourable conditions (a substantial noise level), this may result in an error in determining the time position of the pulse.

Summary

- Modulation is a process whereby the signal spectrum is translated from the baseband into a radio-frequency range.
- In amplitude modulation, the signal envelope is proportional to the instantaneous magnitude of the baseband modulating wave.
- The spectrum of an AM signal is formed by the carrier wave and two symmetrical sidebands.
- The bandwidth required for AM transmission is equal to twice the highest significant frequency in the spectrum of the modulating wave.
- It is possible to modulate the carrier so as to produce a double-sideband (DSB) suppressed-carrier signal and also a single-sideband (SSB) signal.
- In angle modulation, the phase angle of the carrier varies in time in accordance with the message signal.
- The particular forms of angle modulation are frequency modulation and phase modulation.
- The principal parameter of the modulated signal in PM and FM is the angle modulation index equal to the phase deviation.
- Theoretically, there is no bound on the bandwidth of an angle-modulated signal.
- At small values of the modulation index, the bandwidth of the signal is practically equal to twice the maximum frequency of the modulating wave.
- At large values of the modulation index, the bandwidth occupied by the signal is equal to twice the frequency deviation.
- In linear frequency modulation, signals have a practically uniform spectrum within a limited bandwidth, provided the frequency deviation-pulse duration product of the signal is sufficiently large.
- The autocorrelation function of an LFM signal has a lobed structure; the width of the main lobe decreases with the increase in the frequency deviation over the pulse.

Review Questions

1. Name the parameters adopted to define the degree of amplitude modulation.
2. Explain the cause of the distortion occurring in the case of overmodulation.
3. Define the cause of power redistribution in the spectrum of a single-tone AM signal.
4. How are the frequencies of the carrier wave and of the modulating wave related to each other?
5. Define the concept of the rotating phasor used to represent a single-tone AM signal.
6. What is the fundamental difference between the waveforms of a DSB signal and of a conventional AM signal?
7. Why is that the direct demodulation of an SSB signal results in the distortion of the transmitted signal?
8. Outline the similarities and differences between FM and PM signals.
9. State the relationship between the modulating frequency, modulation index, and frequency deviation.
10. What is the spectral composition of FM and PM signals at low values of the modulation index?

11. Explain the difference between the spectra of an AM signal and an FM wave at low values of the modulation index.
12. Which property of Bessel functions indicates that angle-modulated signals occupy a limited bandwidth?
13. How should the angle modulation index be chosen for the carrier to be absent from the modulated signal?
14. What is special about the spectra of FM and PM signals in the case of a nonharmonic modulating wave?
15. Draw the waveform of a rectangular LFM pulse.
16. Explain the physical principle underlying the time compression of an LFM pulse.
17. Draw an approximate plot relating the magnitude of the spectrum of an LFM wave to frequency.
18. How is the frequency deviation-pulse duration product defined for LFM signals?
19. Define the principal characteristics of the phase spectrum of an LFM signal.
20. What is the shape of the autocorrelation function for an LFM signal with a rectangular envelope?
21. What is the limitation of LFM signals from the view-point of the structure of their autocorrelation function?

Problems

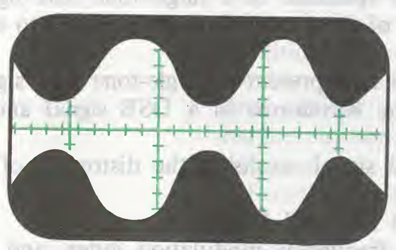
1. An AM wave is defined by

$$u(t) = 130 \times [1 + 0.25\cos(10^2 t + 30°) + 0.75\cos(3 \times 10^2 t + 45°)] \times \cos(10^5 t + 60°)$$

Find the amplitudes and initial phases of all the spectral components and draw the spectral diagram of the signal.

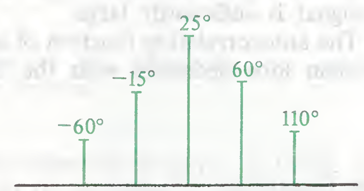
2. Draw a phasor diagram for the signal of Problem 1. Do this for time $t = 0$.

3. The accompanying figure shows the waveform of a single-tone AM signal as displayed by a CRT oscilloscope. Suggest an experimental procedure for determining the modulation depth M from the waveform.



Hint. Observe the instantaneous amplitudes of the signal at the extremal points.

4. Using the waveform of the AM signal shown in the figure find the initial phases of each of the components of the modulating wave.



5. A load resistor of 75 Ω carries an amplitude-modulated current (A) defined by

$$i(t) = 200(1 + 0.8\cos 4 \times 10^3 t)\cos 6 \times 10^6 t$$

Find: (a) the peak power delivered by the source; (b) the mean power dissipated in the load; and (c) the fraction of the total power concentrated in the carrier.

6. An FM signal of 2.7 V amplitude has an instantaneous frequency varying as

$$\omega(t) = 10^9(1 + 10^{-4}\cos 2 \times 10^3 t)$$

Find the modulation index and write a mathematical model of the signal.

7. Determine the angle modulation index for an FM signal, if the modulating frequency is $F = 7$ kHz, the carrier frequency is $f_0 = 180$ MHz, and the peak value of frequency is $f_{max} = 182.5$ MHz.

8. Draw a spectral and a phasor diagram for an angle-modulated signal, if the carrier frequency is 45 MHz, the frequency deviation is 0.3 kHz, and the modulating frequency is 4.5 kHz.

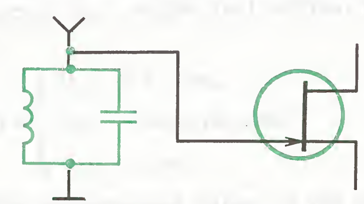
9. There is a single-tone PM signal, and the modulating frequency is $\Omega = 10^4\ s^{-1}$. What should the frequency deviation be for the components at frequencies $\omega_0 \pm \Omega$ to be absent from the signal?

10. Draw a spectral diagram for an FM signal with an amplitude of 35 V and a modulation index of $m = 3$.

11. The output signal from an FM transmitter in the absence of the carrier is 250 V. Measurements show that when a single-tone modulating signal is applied to the transmitter, the output voltage rises to 244 V. Find the frequency modulation index. May it be assumed that narrowband FM is implemented under the conditions stated?

12. A PM radio broadcasting station is operating at a maximum modulation index of 30 (at the maximum volume of the transmitted signal). Assuming that the upper frequency limit of the modulating wave is 16 kHz, find the number of radio channels that can be accommodated within the VHF band (30-300 MHz) without cross-interference.

13. The input circuit of a radio receiver



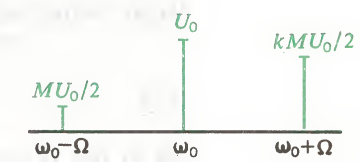
contains a resonant circuit tuned to a carrier frequency of $f_0 = 64$ MHz. The Q-factor of the tuned circuit is 30. Can this circuit be used for the reception of an FM signal whose frequency varies as

$$f(t) = f_0(1 + 0.015\cos 2.8 \times 10^3 t)$$

14. An LFM signal with a rectangular envelope has an amplitude of $U_0 = 10$ V, a duration of $\tau_p = 15$ μs, and a frequency deviation of 40 MHz over the pulse duration. Find: (1) the frequency deviation-pulse duration product of the signal; (2) the quadratic term of the phase spectrum at the boundary of the bandwidth occupied by the signal; and (3) the power spectrum of the signal.

Advanced Problems

15. Analyse how the difference in amplitude between the upper and lower side waves affects the envelope of an AM signal. Consider a single-tone signal with a spectral diagram of the form shown in the accompanying figure,



assuming that $k \neq 1$.

16. Analyse how the failure to observe the constraints imposed on the upper and lower side frequencies affects the envelope of an AM signal. Assume that the two side waves have the same amplitude. Consider a single-tone signal in which the lower side frequency is $\omega_{ls} = \omega_0 - \Omega$, whereas the upper side frequency is $\omega_{us} = \omega_0 + \Omega + \delta$, where $\delta \ll \Omega$.

17. Apply spectral analysis to the envelope of a single-tone SSB signal of the form defined by Eq. (4.18). Propose a numerical estimate for the degree of

distortion in the envelope of the signal.

18. The total phase of a two-tone FM signal varies as

$$\Psi(t) = 2\pi \times 10^8 t + 0.12 \sin 2\pi \times 10^4 t + 0.08 \sin 4\pi \times 10^4 t$$

The amplitude of the unmodulated carrier is $U_0 = 75$ V. Find how the amplitude of the carrier will change upon the application of modulating signals.

19. It is known that one of the methods to improve the correlation properties of LFM signals is to use nonlinear instead of linear modulation. Referring to Chap. 3 in [21], examine the spectrum of such complex signals.

Chapter 5

Band-Limited Signals

As we have learned in Chap. 2, for the complete recovery of a signal from its spectrum it would be necessary to consider all of its components at frequencies from zero to infinity. This is unfeasible for physical reasons. Also, as ω goes to infinity, the contributions from the various spectral components grow negligibly small owing to the properties of the signals whose energy is finite. Finally, any real device designed to transmit and process signals has a finite bandwidth. This becomes especially evident in the case of frequency-selective filters.

In this chapter we will deal with a special class of signals whose spectrum is nonzero only in a frequency interval of finite width. They are aptly called *band-limited.*

5.1 Some Mathematical Models and Properties of Band-Limited Signals

Let D be a finite frequency interval. Then the spectrum of a band-limited signal will be

$S(\omega) \neq 0$ if $\omega \in D$

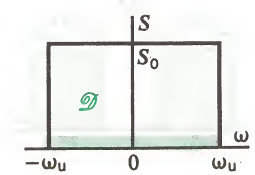
$S(\omega) = 0$ for any other values of frequency

The most general mathematical model for a band-limited signal can be derived from the Fourier inversion formula:

$$s(t) = \frac{1}{2\pi} \int_D S(\omega) \exp(j\omega t)\, d\omega \tag{5.1}$$

Depending on the choice of the frequency interval D and of the spectrum $S(\omega)$, we can obtain a large variety of band-limited signals.

The ideal low-pass signal. Consider a wave whose spectrum has a constant real term in a frequency interval limited to some upper frequency ω_u; outside that interval the spectrum of the signal is zero



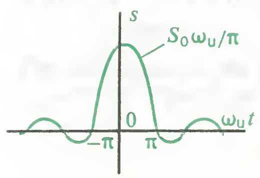
$$S(\omega) = \begin{cases} 0, & \omega < -\omega_u \\ S_0, & -\omega_u \leqslant \omega_u \leqslant \omega_u \\ 0, & \omega > \omega_u \end{cases} \tag{5.2}$$

The instantaneous values of the signal are

$$s(t) = (S_0/2\pi) \int_{-\omega_u}^{\omega_u} \exp(j\omega t)\, d\omega = \frac{S_0 \omega_u}{\pi} \frac{\sin \omega_u t}{\omega_u t} \tag{5.3}$$

The ideal low-pass signal

Let us call this wave the *ideal low-pass signal.* Thus we underscore the fact that it has the simplest spectrum among all the other signals of this kind.



The plot of the ideal low-pass signal constructed on the basis of Eq. (5.3) yields an oscillating curve which is even about the time origin. As the upper frequency limit of the spectrum is raised, the central peak and the frequency of oscillation increase too.

A more general ideal low-pass signal can be obtained by assuming in Eq. (5.2) that the phase of the spectrum is a linear function of frequency

Solve Problem 1

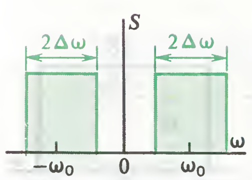
$$S(\omega) = \begin{cases} 0, & \omega < -\omega_u \\ S_0 \exp(-j\omega t_0), & -\omega_u \leqslant \omega \leqslant \omega_u \\ 0, & \omega > \omega_u \end{cases} \tag{5.4}$$

Here t_0 is a parameter defining the time shift of the wave such that the spectrum in (5.4) corresponds to the low-pass signal defined by

$$s(t) = \frac{S_0\omega_u}{\pi} \frac{\sin \omega_u(t - t_0)}{\omega_u(t - t_0)}$$

The ideal low-pass signal is an idealized representation of the response of a low-pass filter excited by a wave whose spectrum is the same over all frequencies, which is recognized as the delta-impulse. In this representation, it is assumed that the frequency response of the low-pass filter is approximated with sufficient accuracy by a rectangular function having the specified upper frequency limit.

The ideal bandpass signal. Analyse the mathematical model of a bandpass signal whose spectrum is limited to a frequency band of width $BW = 2\Delta\omega$, centred at $\pm\omega_0$. If within that band the spectrum of the signal is constant:



$$S(\omega) = \begin{cases} S_0, & -\omega_0 - \Delta\omega < \omega < -\omega_0 + \Delta\omega \\ & \omega_0 - \Delta\omega < \omega < \omega_0 + \Delta\omega \\ 0, & \text{outside the band} \end{cases} \tag{5.5}$$

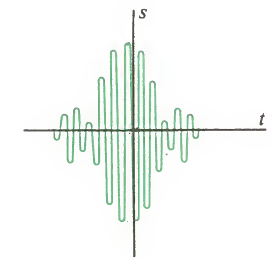
then, by analogy with the previous case, we shall call this signal the *ideal bandpass signal.*

The ideal bandpass signal

The instantaneous values of an ideal bandpass signal can be found, using the inverse Fourier transform:

$$s(t) = (S_0/\pi) \int_{\omega_0 - \Delta\omega}^{\omega_0 + \Delta\omega} \cos \omega t \, d\omega = \frac{2S_0\Delta\omega}{\pi} \frac{\sin \Delta\omega t}{\Delta\omega t} \cos \omega_0 t \tag{5.6}$$

The structure of an ideal bandpass signal is illustrated in the accompanying plot. As is seen, the r.f. oscillations at ω_0 are accompanied by time variations in their instantaneous amplitudes.



The function $\sin(\Delta\omega t)/(\Delta\omega t)$ represents the slowly varying envelope of the ideal bandpass signal to within the scale factor $2S_0\Delta\omega/\pi$. If $\Delta\omega/\omega_0 \ll 1$, that is, if the relative bandwidth of the ideal bandpass signal is substantially less than unity, the envelope varies much more slowly than the r.f. carrier.

A method, at least theoretical, to generate an ideal bandpass signal is obvious: We should apply a broadband excitation such as the delta-impulse to the input of an ideal bandpass filter which transmits only frequencies within the interval $[\omega_0 - \Delta\omega, \omega_0 + \Delta\omega]$.

Estimating some parameters of a band-limited signal. From an analysis of the equations derived in the previous section, it would be seen that the peak values of the signals involved are directly proportional to their bandwidth. On the other hand, as the bandwidth of a signal is reduced, the signal (or its envelope) varies with time progressively more slowly.

Work Problem 2

It is an easy matter to develop numerical characteristics which would define the limiting values for the energy, amplitude and derivative of a band-limited signal in general form.

Let, for example, $s(t)$ be a signal whose spectrum is nonzero only within the frequency band $-\omega_u \leqslant \omega \leqslant \omega_u$. Then the following estimate of the signal energy is obvious:

$$E_s = \frac{1}{2\pi} \int_{-\omega_u}^{\omega_u} S(\omega) S^*(\omega) \, d\omega \leqslant (\omega_u/\pi) S_{max}^2 \tag{5.7}$$

where S_{max} is the maximum magnitude of the spectrum.

Let us find the limiting value of the signal amplitude. By the inverse Fourier transform

$$s(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} S(\omega) \exp(j\omega t) \, d\omega$$

the maximum value of a signal is limited from above to a value equal to the sum of the magnitudes of all complex amplitudes corresponding to elementary intervals on the frequency axis. Since, on the basis of Eq. (2.17), the frequency interval of width $\Delta\omega$, forming the neighbourhood around the chosen frequency, is characterized by a complex amplitude

$$\Delta A_\omega = \frac{1}{\pi} S(\omega) \Delta\omega$$

it follows that the instantaneous value of the signal in question satisfies the inequality

$$s_{max} \leqslant (\omega_u/\pi) S_{max} \tag{5.8}$$

The "=" sign applies only when the signal is of the ideal low-pass

type: Its spectrum must be constant within the bandwidth and also all spectral components must be able to combine coherently at some instant of time.

Now it is easy to develop an estimate of the time derivative. We take the limit of the relation

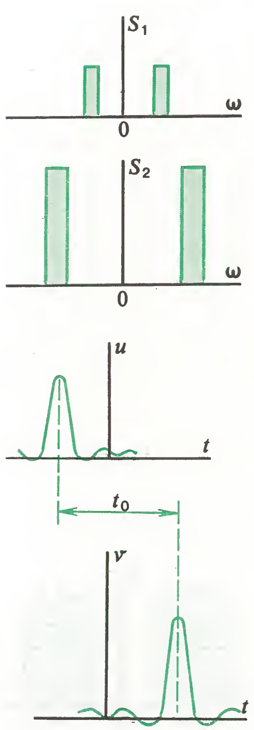
$$\lim_{\Delta t \to 0} \frac{s(t+\Delta t)-s(t)}{\Delta t} = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \int_{-\omega_u}^{\omega_u} S(\omega)\,[\exp(j\omega\Delta t)-1] \times \exp(j\omega t)\,d\omega$$

and note that

$$\exp(j\omega\Delta t) - 1 \approx j\omega\Delta t$$

for Δt going to zero. By analogy with the previous case, we obtain the following inequality

$$|\,ds/dt\,|_{\max} \leqslant (\omega_u^2/\pi)\,S_{\max} \tag{5.9}$$



Orthogonal band-limited signals. The constraints imposed on the bandwidth of a signal permit us to find interesting and important classes of orthogonal signals. The simplest example is the orthogonality of two bandpass signals whose spectra exist in non-intersecting domains. That the scalar product of these signals is equal to zero stems directly from the generalized Rayleigh formula.

The orthogonality is less obvious in the case of band-limited signals shifted in time relative to one another.

Consider two ideal baseband signals $u(t)$ and $v(t)$. Their S_0 and ω_u are the same, but the signal $v(t)$ lags behind the signal $u(t)$ by a time t_0 such that its spectrum is defined by

$$S_v(\omega) = S_u(\omega)\exp(-j\omega t_0)$$

The scalar product of the two signals is

$$(u, v) = \frac{S_0^2}{2\pi}\int_{-\omega_u}^{\omega_u} \exp(j\omega t_0)\,d\omega = \frac{S_0^2\omega_u}{\pi}\,\frac{\sin\omega_u t_0}{\omega_u t_0} \tag{5.10}$$

As follows from Eq. (5.10), two ideal baseband signals identical in waveform are orthogonal if the time shift between them satisfies the condition

$$\omega_u t_0 = k\pi, \quad k = \pm 1, \ \pm 2, \ \ldots \tag{5.11}$$

The minimum time shift just sufficient for the signals to be orthogonal occurs at $k = \pm 1$ and is equal to

$$t_0 = \pm(\pi/\omega_u) = \pm 1/2f_u \tag{5.12}$$

A fact of fundamental importance is not only that two signals are made orthogonal. We have formed an infinite orthogonal basis which can serve as a coordinate system for the expansion of an arbitrary signal whose spectrum does not contain any frequencies above ω_u.

Plots of the signals involved appear in Fig. 5.1.

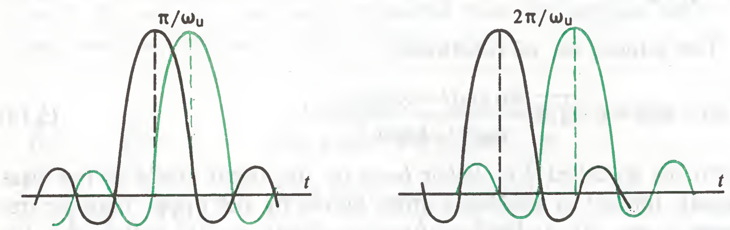


Fig. 5.1 Plots of two ideal baseband signals shifted in time for $t_0 = \pi/\omega_h$ and $t_0 = 2\pi/\omega_h$

Another important fact is this: When one of the signals reaches an absolute maximum, all the other signals in the set cross zero.

5.2 The Kotelnikov Theorem

This theorem*, proved by Kotelnikov in 1933, is one of the fundamental postulates of communication theory. It provides for the recovery of the instantaneous values of a band-limited signal to any desired accuracy from its samples taken at equal time intervals.

Academician V. A. Kotelnikov is a prominent Soviet scientist in the field of telecommunications and radio physics

Construction of an orthonormal basis. As has already been shown, any two band-limited signals belonging to the set

$$u_k(t) = A\,\frac{\sin\omega_u(t - k\pi/\omega_u)}{\omega_u(t - k\pi/\omega_u)}, \quad k = 0, \ \pm 1, \ \pm 2 \ldots \tag{5.13}$$

are orthogonal. Through the proper choice of the amplitude coefficient A we can make the norm of each of these signals equal to unity. The result will be an orthonormal basis for the expansion of an arbitrary band-limited signal into a generalized Fourier series.

We shall limit ourselves to the function

$$u_0(t) = A\,\frac{\sin\omega_u t}{\omega_u t}$$

because the norm of any signal u_k is the same, irrespective of the

* It is usually called the *sampling theorem* outside the Soviet Union. – *Translator's note.*

time shift. Since

$$\|u_0\|^2 = (A^2/\omega_u^2)\int_{-\infty}^{\infty}\frac{\sin^2\omega_u t}{t^2}\,dt = \pi A^2/\omega_u$$

then the functions u_k will be orthonormal if

$$A = \sqrt{\omega_u/\pi}$$

The infinite set of functions

$$Sc_k(t;\,\omega_u) = \sqrt{\omega_u/\pi}\,\frac{\sin\omega_u(t - k\pi/\omega_u)}{\omega_u(t - k\pi/\omega_u)} \tag{5.14}$$

The Kotelnikov basis

form the so-called *Kotelnikov basis* on the linear space of low-pass signals limited in frequency from above by the upper limiting frequency, ω_u. An individual function $Sc_k(t;\,\omega_u)$ is called the *kth sample function.*

The Kotelnikov series. If $s(t)$ is an arbitrary signal whose spectrum is nonzero only within the frequency band $-\omega_u < \omega < < \omega_u$, it can be expanded into a generalized Fourier series in terms of the Kotelnikov basis, that is, represented as

$$s(t) = \sum_{k=-\infty}^{\infty} c_k Sc_k(t,\,\omega_u) \tag{5.15}$$

The coefficients of the Kotelnikov series are the scalar products of the signal being expanded and the kth sample function:

$$c_k = [s(t),\, Sc_k(t,\,\omega_u)] \tag{5.16}$$

These coefficients can conveniently be determined by using the generalized Rayleigh formula. It is easy to verify that within the frequency interval $(-\omega_u \leqslant \omega \leqslant \omega_u)$ the kth sample function has a spectrum equal to $\sqrt{\pi/\omega_u}\exp(-j\omega k\pi/\omega_u)$. It is obvious from a comparison of Eqs. (5.3) and (5.14). Then, if $S(\omega)$ is the spectrum of the transmitted signal $s(t)$, we find that

$$c_k = \sqrt{\pi/\omega_u}\left[\frac{1}{2\pi}\int_{-\omega_u}^{\omega_u} S(\omega)\exp(jk\pi\omega/\omega_u)\,d\omega\right] \tag{5.17}$$

The expression in the square brackets is nothing but the instantaneous value of $s(t_k) = s_k$ at the kth sampling point:

$$t_k = k\pi/\omega_u = k/2f_u$$

Thus,

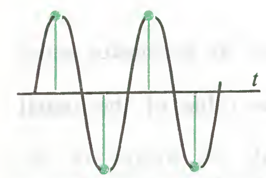
$$c_k = \sqrt{\pi/\omega_u}\,s_k$$

Hence, the final expression for the Kotelnikov series is

$$s(t) = \sum_{k=-\infty}^{\infty} s_k\,\frac{\sin\omega_u(t - k\pi/\omega_u)}{\omega_u(t - k\pi/\omega_u)} \tag{5.18}$$

The statement of the Kotelnikov theorem

On the basis of Eq. (5.18), the Kotelnikov theorem may be stated as follows: *An arbitrary signal band-limited to frequencies not higher than f_u Hz can be completely recovered from its samples taken at equal sampling intervals $1/(2f_u)$ seconds apart.*

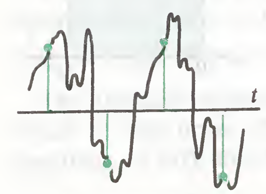


Example 5.1. *Let there be a signal $s(t) = \cos\omega_u t$. The samples of* the signal are defined by

$$s_k = \cos(\omega_u t_k) = \cos(k\pi) = (-1)^k$$

Hence, it is seen that the harmonic waveform at the highest frequency allowed by the Kotelnikov theorem is represented by two samples per period. If the initial phase of the signal is zero, then the samples are equal in magnitude, but their signs alternate. The Kotelnikov theorem permits us to write the following representation of the harmonic signal, which does not appear obvious at first glance:

$$\cos\omega t = \sum_{k=-\infty}^{\infty}\left\{\frac{\sin\omega(t - 2k\pi/\omega)}{\omega(t - 2k\pi/\omega)} - \frac{\sin\omega[t - (2k+1)\pi/\omega]}{\omega[t - (2k+1)\pi/\omega]}\right\}$$



If, on the other hand, the requirement of the Kotelnikov theorem is not observed and samples are taken not frequently enough, the recovery of the signal in accord with Eq. (5.18) is accompanied by an error which may be considerable. For example, the samples may be $s_k = (-1)^k$, as in the previous case, but the signal whose spectrum contains frequencies higher than ω_u will have a more complex structure than a harmonic wave.

Hardware for the synthesis of signals represented by the Kotelnikov series. A valuable advantage of the Kotelnikov theorem is its constructive character. It not only states the fact that a signal can be expanded into a suitable series, but also defines the method by which a continuous signal can be recovered from its samples (Fig. 5.2).

Solve Problem 6

Let there be a set of generators producing sample functions $Sc_k(t;\,\omega_u)$ across their output terminals. The generators are of the controlled type – the amplitudes of their output signals are proportional to the samples s_k. If, now, we combine the output waves by applying them to an adder, the adder output will, in

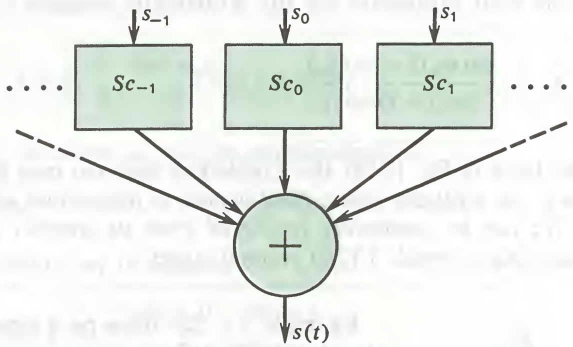
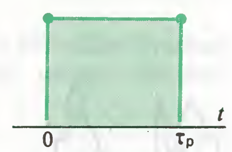


Fig. 5.2 Set-up for the synthesis of a signal from its Kotelnikov series

accord with Eq. (5.18), be the instantaneous value of the signal being synthesized, $s(t)$.

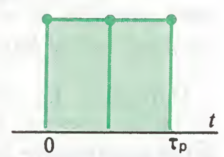
An approximate representation of signals by means of the Kotelnikov series. Cases may arise in which it is known in advance that the greater proportion of the energy of a signal having an unbounded bandwidth is concentrated in the low-frequency portion of the spectrum. Such signals can approximately be represented by the Kotelnikov series.

Example 5.2. *A rectangular video pulse of unit amplitude and of τ_p duration is not band-limited, yet the magnitude of its spectrum decreases with increasing frequency at a sufficiently high rate (as $1/\omega$).*



If we represent the signal with two samples one taken at the start and the other at the finish of the pulse, its spectrum will include all the components up to the highest frequency $\omega_u = \pi/\tau_p$. Using Eq. (5.18), we find an approximate mathematical model for the signal:

$$s(t) = \frac{\sin(\pi t/\tau_p)}{\pi t/\tau_p} + \frac{\sin(\pi/\tau_p)(t-\tau_p)}{(\pi/\tau_p)(t-\tau_p)} \tag{5.19}$$



If we represent the same signal with three equidistant samples, we will, as can readily be seen, include all frequencies up to $\omega_u = 2\pi/\tau_p$. Therefore,

$$s(t) = \frac{\sin(2\pi t/\tau_p)}{2\pi t/\tau_p} + \frac{\sin(2\pi/\tau_p)(t-\tau_p/2)}{(2\pi/\tau_p)(t-\tau_p/2)} + \frac{\sin(2\pi/\tau_p)(t-\tau_p)}{(2\pi/\tau_p)(t-\tau_p)} \tag{5.20}$$

The corresponding plots appear in Fig. 5.3.

Naturally, the accuracy of the approximation will improve as we

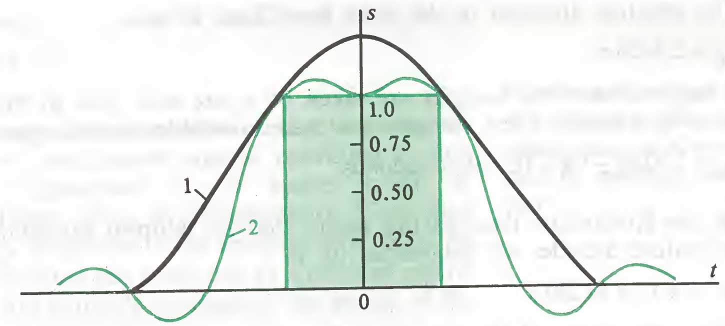


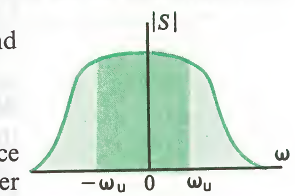
Fig. 5.3 Approximation of a video pulse: (*1*) with two members of the Kotelnikov series; (*2*) with three members of the Kotelnikov series

take more samples, that is, as we reduce the time spacing between the samples.

The error in the Kotelnikov series approximation of an arbitrary signal. If $s(t)$ is an arbitrary signal, it may be represented as a sum

$$s(t) = s_{bl}(t, \omega_u) + s_e(t)$$

where $s_{bl}(t, \omega_u)$ is a signal band-limited to the highest frequency ω_u, and $s_e(t)$ is the approximation error signal occupying an unlimited frequency band $\omega > \omega_u$.



Their spectra do not overlap, so s_{bl} and s_e are orthogonal, and their powers, that is, their norms squared, combine:

$$\|s\|^2 = \|s_{bl}\|^2 + \|s_e\|^2$$

The absolute measure of the approximation error is the distance equal to the norm of the error signal. If $W_s(\omega)$ is the power spectrum of $s(t)$, then by the Rayleigh theorem

$$\|s_e\| = \left[\left(\frac{1}{\pi}\right)\int_{\omega_u}^{\infty} W_s(\omega)\,d\omega\right]^{1/2} \tag{5.21}$$

Example 5.3. *There is an exponential video pulse* $s(t) = \exp(-\alpha t)\sigma(t)$, *for which the power spectrum is*

$$W_s = 1/(\alpha^2 + \omega^2)$$

and the power is

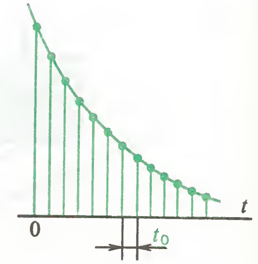
$$\|s\|^2 = \frac{1}{\pi}\int_0^{\infty} \frac{d\omega}{(\alpha^2+\omega^2)} = \frac{1}{2\alpha}$$

The effective duration of the pulse (see Chap. 2) is

$$\tau_p = 2.3026/\alpha$$

Suppose that the samples are taken at a rate such that in the time τ_p a total of ten samples are made available, spaced apart

$$t_0 = 2.3026 \div 9\alpha = 0.2558/\alpha \text{ seconds}$$



By the Kotelnikov theorem this means that the adopted sampling procedure includes all frequencies up to

$$\omega_u = \pi/t_0 = 12.281\alpha$$

Hence the norm of the error signal is

$$\|s_e\| = \left[\frac{1}{\pi}\int_{\omega_u}^{\infty}\frac{d\omega}{\alpha^2+\omega^2}\right]^{1/2} = \left[\frac{1}{\alpha\pi}\left(\frac{\pi}{2}-\arctan\frac{\omega_u}{\alpha}\right)\right]^{1/2} = 0.1608/\sqrt{\alpha}$$

Here the theoretical minimum of the error is meant

Since

$$\|s\| = 0.7071/\sqrt{\alpha}$$

it follows that the relative approximation error is

$$\|s_e\|/\|s\| = 0.1608 \div 0.7071 = 0.2274$$

Work Problem 7

As is seen, the adopted sampling rate is not high enough for the approximation to be sufficiently accurate.

Dimensionality of the space of band- and duration-limited signals. As follows from the examples given in Chap. 2 to demonstrate the calculation of the spectrum of pulse signals, theoretically any signal of finite duration has a spectrum extending along the frequency axis without bound. In mathematics, this statement is proved rigorously and in the general form.

For practical purposes, however, it is customary to consider idealized models of signals limited in duration as well as in bandwidth. Such models can fairly accurately represent the signals existing in real communication channels.

Let T be the duration of such a signal, and f_u be its upper frequency limit or, which is the same, its bandwidth in hertz. Then the bandwidth-duration product of the signal (see Chap. 4) will be

$$BW = Tf_u$$

For this signal to be described completely (albeit approximately for the reasons just stated), we must take a number N of independent samples, such that $N = T/t_0 = 2Tf_u$. This expression defines the dimensionality of the space of band- and duration-limited signals.

As a rule, the number N is fairly large. For example, in order to represent a radio broadcasting signal band-limited to a frequency of 12 kHz and duration-limited to 1 min, we shall need

$$2 \times 60 \times 1.2 \times 10^4 = 1.44 \times 10^6$$

independent numbers.

In information theory, the dimensionality of signal space is a measure of message content

In his time, Claude Shannon proposed a convenient technique for interpreting duration- and band-limited signals, consisting in that each such signal is represented by a single point in a finite-dimensional Euclidean space of dimensionality $2Tf_u$. Now the sample s_k appears as a projection of the representative point on the kth coordinate axis. Because the space has an Euclidean metric and the coordinate axes are mutually orthogonal, the length of the signal vector is

$$r_s = \left(\sum_{k=1}^{2Tf_u} s_k^2\right)^{1/2} \tag{5.22}$$

The quantity r_s can be expressed in terms of the signal energy E_s in the following way. Since

$$E_s = \sum_{k=1}^{2Tf_u} c_k^2 = \frac{1}{2f_u}\sum_{k=1}^{2Tf_u} s_k^2$$

then it follows that

$$r_s = \sqrt{2E_s f_u} = \sqrt{2Tf_u P_m} \tag{5.23}$$

where P_m is the mean power of the signal. Hence, any signals with fixed parameters T and f_u, the mean power of which does not exceed P_0, are represented by points lying within a multidimensional sphere of radius

$$\rho(P_0) = \sqrt{2Tf_u P_0} \tag{5.24}$$

It should be noted in conclusion that the statement of the Kotelnikov theorem may be somewhat extended [18]. It is true that for a finite-dimensional representation of the signal type in question we need $2Tf_u$ samples. It is not mandatory, however, to require that the samples be equidistant in time. Also, we may halve the sampling rate, but at each point we should measure both the instantaneous value of the signal and its first derivative.

5.3 Narrowband Signals

It is assumed that $\omega_0 \neq 0$

This section will be concerned with a special class of band-limited signals which appear at the output of frequency-selective circuits. By definition, a signal is called narrowband if its spectrum is non-zero only within the frequency bands of width BW, in the neighbourhood of points $\pm\omega_0$, such that $BW/\omega_0 \ll 1$.

The reference frequency

As a rule, the frequency ω_0, called the *reference frequency* of the

signal, may be taken as being the same as the centre frequency of the spectrum. In the general case, however, its choice is sufficiently arbitrary.

The mathematical model of a narrowband signal. A straightforward approach to forming a mathematical model for a narrowband signal is as follows. As will be recalled (see Chap. 2), if $f_1(t)$ is a low-pass signal whose spectrum is concentrated in the neighbourhood of ω_0, then the wave

$$s_1(t) = f_1(t) \cos \omega_0 t$$

will possess all the necessary properties of a narrowband signal because its spectrum will be concentrated near the points $\pm\omega_0$. The same is true of the signal

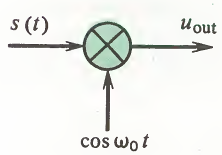
$$s_2(t) = f_2(t) \sin \omega_0 t$$

which only differs in the phase of the "fast" term. The most general expression for the mathematical model of a narrowband signal can be derived by analysing the following linear combination

The in-phase and quadrature amplitudes

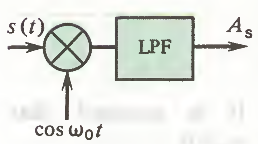
$$s(t) = A_s(t) \cos \omega_0 t - B_s(t) \sin \omega_0 t \tag{5.25}$$

The two time functions, $A_s(t)$ and $B_s(t)$, are of low frequency character in the sense that they change only slightly over a period of the high-frequency wave, $T = 2\pi/\omega_0$. Customarily, the function $A_s(t)$ is called the *in-phase amplitude* of a narrowband signal, $s(t)$, for a specified reference frequency ω_0, and the function $B_s(t)$ is called its *quadrature amplitude*.



Experimentally, the instantaneous values of $A_s(t)$ and $B_s(t)$ can be determined as follows. If we apply the narrowband signal $s(t)$ to one input of a multiplier (mixer) and an auxiliary wave varying in time as $\cos \omega_0 t$, to the other input, the output will be the signal

$$u_{out}(t) = A_s(t) \cos^2 \omega_0 t - \frac{1}{2} B_s(t) \sin 2\omega_0 t$$
$$= A_s(t)/2 + [A_s(t)/2] \cos 2\omega_0 t - [B_s(t)/2] \sin 2\omega_0 t \tag{5.26}$$



Now let the output signal be applied to a low-pass filter, LPF, which suppresses frequencies of the order of $2\omega_0$. Obviously, the output of the low-pass filter will be a low-frequency wave proportional to the in-phase amplitude A_s. The quadrature amplitude $B_s(t)$ of the narrowband signal $s(t)$ can be extracted if we apply another auxiliary wave $\sin \omega_0 t$ to the second input of the multiplier.

The complex representation of narrowband signals. It is a widely accepted fact that many problems in the theory of linear circuits

can be tackled by the method of complex amplitudes. By this method, a harmonic wave is represented as the real or the imaginary part of a complex function:

$$U_0 \cos(\omega_0 t + \varphi_0) = \mathrm{Re}\,[U_0 \exp(j\varphi_0) \exp(j\omega_0 t)]$$

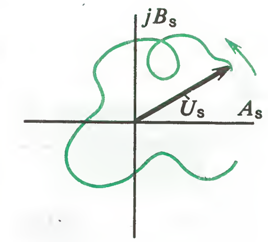
$$U_0 \sin(\omega_0 t + \varphi_0) = \mathrm{Im}\,[U_0 \exp(j\varphi_0) \exp(j\omega_0 t)]$$

The time-invariant number $\dot{U} = U_0 \exp(j\varphi_0)$ is called the *complex amplitude* of the harmonic wave.

From a physical point of view, narrowband signals are *quasi-harmonic waves*. Therefore, it is natural to try to generalize the method of complex amplitudes to signals of the type (5.25). Let us introduce a complex low-frequency function

The complex envelope

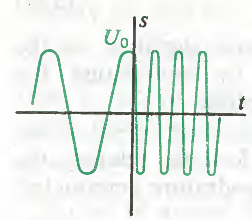
$$\tilde{U}_s(t) = A_s(t) + jB_s(t) \tag{5.27}$$



called the *complex envelope of a narrowband signal*. It is easy to verify that

$$s(t) = A_s \cos \omega_0 t - B_s \sin \omega_0 t = \mathrm{Re}\,[\tilde{U}_s(t) \exp(j\omega_0 t)] \tag{5.28}$$

Thus, for a narrowband signal, the complex envelope is the same as the complex amplitude for a simple harmonic wave. However, the complex envelope is, in the general case, time-dependent, because the phasor $\tilde{U}_s(t)$, as it moves about in the complex plane, varies in both magnitude and position.



Example 5.4. *A narrowband signal $s(t)$ is a harmonic wave at $t<0$ and $t>0$, but at $t=0$ its frequency undergoes a step change:*

$$s(t) = \begin{cases} U_0 \cos \omega_0 t, & t<0 \\ U_0 \cos \omega_1 t, & t>0 \end{cases}$$

Taking ω_0 as the reference frequency, we obtain the following expression for the complex envelope of the signal:

$$\tilde{U}_s(t) = \begin{cases} U_0, & t<0 \\ U_0 \exp[j(\omega_1 - \omega_0)t], & t>0 \end{cases}$$

It is important to stress that *the choice of the reference frequency is fairly arbitrary* and dictated by mathematical convenience. For example, if the reference frequency is chosen to be $(\omega_0 + \omega_1)/2$, the complex envelope of the signal in question will have the form

$$\tilde{U}_s(t) = \begin{cases} U_0 \exp\left(j\,\dfrac{\omega_0 - \omega_1}{2}\,t\right), & t<0 \\ U_0 \exp\left(j\,\dfrac{\omega_1 - \omega_0}{2}\,t\right), & t>0 \end{cases}$$

The physical envelope, the total phase, and the instantaneous frequency. Equation (5.27) defining the complex envelope may be re-cast in exponential form:

$$\tilde{U}_s(t) = U_s(t)\exp\left[j\varphi_s(t)\right] \tag{5.29}$$

In dealing with modulated signals we have been using precisely this concept of the envelope

Here, $U_s(t)$ is a real, nonnegative function of time, called the *physical envelope* (or, simply, the *envelope*); $\varphi_s(t)$ is the slowly varying initial phase of the narrowband signal. The two quantities are related to the in-phase and quadrature amplitudes:

$$A_s(t) = U_s(t)\cos\varphi_s(t) \tag{5.30}$$

$$B_s(t) = U_s(t)\sin\varphi_s(t)$$

Hence, we derive one more useful form for the mathematical model of a narrowband signal

$$s(t) = U_s(t)\cos\left[\omega_0 t + \varphi_s(t)\right] \tag{5.31}$$

▲ Solve Problems 8-10

As we have already done in connection with angle modulation, we introduce the *total phase* of a narrowband signal

$$\psi_s(t) = \omega_0 t + \varphi_s(t)$$

and define the *instantaneous frequency* of the signal as the time derivative of the total phase:

$$\omega_s(t) = \omega_0 + \mathrm{d}\varphi_s/\mathrm{d}t \tag{5.32}$$

In accordance with Eq. (5.31), a narrowband signal is, in the general case, a complex wave produced by modulating the harmonic carrier both in amplitude and in phase.

The physical envelope of a narrowband signal and its properties. On the basis of Eq. (5.30), let us write a formula relating the physical envelope $U_s(t)$ to the in-phase and quadrature amplitudes:

$$U_s(t) = \sqrt{A_s^2 + B_s^2} \tag{5.33}$$

(The arithmetic value of the root is taken.)

As already noted, there is more than one way to define the complex envelope of a narrowband signal. If, instead of the frequency ω_0 appearing in Eq. (5.28), we take some other frequency, $\omega_0' = \omega_0 + \Delta\omega$, then the signal $s(t)$ will be defined by

$$s(t) = \mathrm{Re}\left[\tilde{U}_s(t)\exp(-j\Delta\omega t)\exp(j\omega_0' t)\right]$$

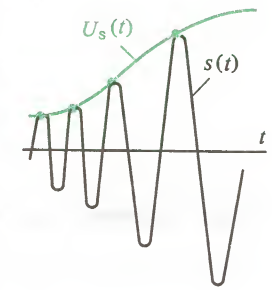
and its complex envelope will take the form

$$\tilde{U}_s'(t) = \tilde{U}_s(t)\exp(-j\Delta\omega t) \tag{5.34}$$

Meanwhile, however, the physical envelope, which is the magnitude of the complex envelope, will remain unchanged, because the term $\exp(-j\Delta\omega t)$ has a magnitude of unity.

Another important property of the physical envelope consists in that at each instant of time

$$|s(t)| \leqslant U_s(t)$$



The envelope of a narrow-band signal

The validity of the statement stems directly from Eq. (5.31). The "=" sign applies at the instants when

$$\cos\left[\omega_0 t + \varphi_s(t)\right] = 1$$

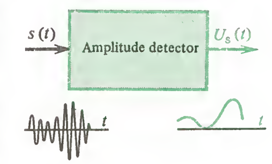
But the derivatives of the signal and of its envelope are the same:

$$s'(t) = U_s'(t)\cos\left[\omega_0 t + \varphi_s(t)\right]$$

$$-\left[\omega_0 + \varphi_s'(t)\right]U_s(t)\sin\left[\omega_0 t + \varphi_s(t)\right]$$

Therefore, the physical envelope does "enclose" the narrowband signal and has the meaning of its instantaneous amplitude.

The concept of the envelope is of important significance to telecommunications because in practice wide use is made of special devices, called *amplitude detectors (demodulators)*, which are capable of re-creating the envelope of a narrowband signal with high fidelity.



Properties of the instantaneous frequency of a narrowband signal. If the complex envelope of a narrowband signal is represented by a phasor which rotates in the complex plane at a constant angular velocity Ω, that is,

$$\tilde{U}_s(t) = U_s(t)\exp(\pm j\Omega t)$$

then, in accord with Eq. (5.32), the instantaneous frequency of the signal must be constant in time:

$$\omega_s = \omega_0 \pm \Omega$$

▲ Work Problem 11

It may be argued that this signal is a harmonic wave modulated only in amplitude, but not in phase. Among other things, if one of the amplitudes, A_s or B_s, reduces identically to zero, then the instantaneous frequency at any time will be $\omega_s = \omega_0$.

In the general case, however, the instantaneous frequency varies with time in the following manner:

$$\omega_s(t) = \omega_0 + \frac{\mathrm{d}}{\mathrm{d}t}\arctan(B_s/A_s) = \omega_0 + \frac{B_s' A_s - A_s' B_s}{A_s^2 + B_s^2} \tag{5.35}$$

Relationship between the spectra of a narrowband signal and of its complex envelope. Let $G_s(\omega)$ denote the spectrum of the complex envelope of a narrowband signal, $s(t)$. The spectrum of the

signal itself is

$$S(\omega)=\int_{-\infty}^{\infty}\mathrm{Re}[\tilde{U}_s(t)\exp(j\omega_0 t)]\exp(-j\omega t)\,\mathrm{d}t$$

$$=\frac{1}{2}\int_{-\infty}^{\infty}\tilde{U}_s\exp[-j(\omega-\omega_0)t]\,\mathrm{d}t$$

$$+\frac{1}{2}\int_{-\infty}^{\infty}\tilde{U}_s^*\exp[-j(\omega+\omega_0)t]\,\mathrm{d}t$$

$$=\frac{1}{2}G_s(\omega-\omega_0)+\frac{1}{2}G_s^*(-\omega-\omega_0) \tag{5.36}$$

Thus, the spectrum of a narrowband signal can be found by translating the spectrum of its complex envelope from the neigbourhood of the zeroth frequency into the neighbourhood of points $\pm\omega_0$; the amplitudes of all spectral components are halved, and the negative-frequency spectrum is obtained by taking the complex conjugate of the positive-frequency spectrum.

Using Eq. (5.36), we can find the spectrum of the complex envelope of a narrowband signal from the known spectrum of the signal itself, and then determine the physical envelope and the instantaneous frequency of the signal.

Example 5.5. *Analyse a narrowband signal, $s(t)$, whose spectrum is asymmetrical about the frequency ω_0:*

$$S(\omega)\Big|_{\omega>0}=\begin{cases}(S_0/2)\exp\,[-b(\omega-\omega_0)], & \omega>\omega_0\\ 0, & 0<\omega<\omega_0\end{cases}$$

On the basis of Eq. (5.36), the spectrum of the complex envelope is

$$G_s(\omega)=\begin{cases}S_0\exp(-b\omega), & \omega>0\\ 0, & \omega<0\end{cases}$$

Using the inverse Fourier transform, we get

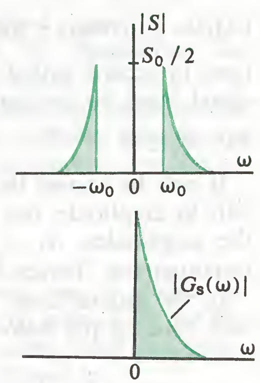
$$\tilde{U}_s(t)=(S_0/2\pi)\int_0^{\infty}\exp(-b+jt)\omega\,\mathrm{d}\omega=(S_0/2\pi)[1/(b-jt)]$$

The in-phase and quadrature amplitudes of the signal respectively are

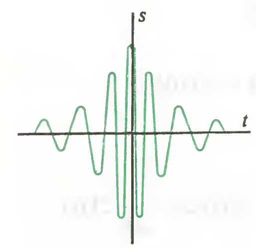
$$A_s(t)=(S_0/2\pi)\,b/(b^2+t^2)$$

and

$$B_s(t)=(S_0/2\pi)\,t/(b^2+t^2)$$



▲ Solve Problem 12



The physical envelope is

$$U_s(t)=|\tilde{U}_s(t)|=(S_0/2\pi)(1/\sqrt{b^2+t^2})$$

The instantaneous frequency of the signal

$$\omega_s(t)=\omega_0+\frac{\mathrm{d}}{\mathrm{d}t}\arctan(t/b)=\omega_0+b/(b^2+t^2)$$

is a maximum, ω_0+1/b, at time $t=0$.

The waveform of the signal in question is a symmetric radio pulse whose carrier varies with time.

5.4 The Analytic Signal and the Hilbert Transform

In this section we will discuss one more method for the complex representation of signals, frequently used in theoretical studies. A distinction of the method is that the concepts of the envelope and of the instantaneous frequency of a signal can be introduced without the degree of uncertainty inherent in the method of complex envelopes.

The analytic signal. Euler's formula

$$\cos\omega t=[\exp(j\omega t)+\exp(-j\omega t)]/2$$

representing a harmonic wave as a sum of two complex conjugate functions suggests that an arbitrary signal $s(t)$ with a known spectrum $S(\omega)$ can be uniquely defined as a sum of two components each of which contains only positive or only negative frequencies:

$$s(t)=\frac{1}{2\pi}\int_{-\infty}^{\infty}S(\omega)\exp(j\omega t)\,\mathrm{d}\omega$$

$$=\frac{1}{2\pi}\int_{-\infty}^{0}S(\omega)\exp(j\omega t)\,\mathrm{d}\omega+\frac{1}{2\pi}\int_{0}^{\infty}S(\omega)\exp(j\omega t)\,\mathrm{d}\omega \tag{5.37}$$

Let the function

$$z_s(t)=\frac{1}{\pi}\int_0^{\infty}S(\omega)\exp(j\omega t)\,\mathrm{d}\omega \tag{5.38}$$

be called the *analytic signal* corresponding to a real signal $s(t)$.

By a change of variables, $\xi=-\omega$, the first integral on the right-

hand side of Eq. (5.37) can be re-arranged as

$$\frac{1}{2\pi}\int_{-\infty}^{0} S(\omega)\exp(j\omega t)\,d\omega = -\frac{1}{2\pi}\int_{\infty}^{0} S(-\xi)\exp(-j\xi t)\,d\xi$$

$$= \frac{1}{2\pi}\int_{0}^{\infty} S(-\xi)\exp(-j\xi t)\,d\xi = \frac{1}{2\pi} z_s^*(t)$$

Therefore, Eq. (5.37) relates the signals $s(t)$ and $z_s(t)$ as

$$s(t) = [z_s(t) + z_s^*(t)]/2$$

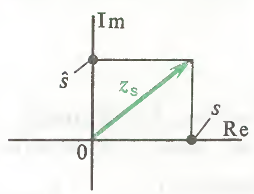
or, which is the same,

$$s(t) = \mathrm{Re}[z_s(t)] \tag{5.39}$$

The imaginary part of the analytic signal

$$\hat{s}(t) = \mathrm{Im}[z_s(t)] \tag{5.40}$$

The conjugate signal



is the conjugate of the original signal $s(t)$.

Thus, the analytic signal

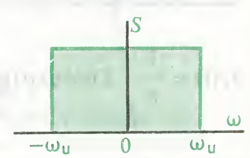
$$z_s(t) = s(t) + j\hat{s}(t) \tag{5.41}$$

is represented in the complex plane by a rotating phasor whose magnitude and phase angle vary with time. The projection of the analytic signal on the real axis at any time is equal to the original signal $s(t)$.

By introducing the concepts of the analytic and the conjugate signals, we cannot derive any new information that is not contained in the mathematical model of the signal $s(t)$. However, these two concepts open a direct road to a systematic approach to the analysis of various signals, especially narrowband waves.

Now we shall demonstrate how the analytic signal can be found from the known spectrum of the original signal.

Example 5.6. *Let $s(t)$ be an ideal low-pass signal of known S_0 and ω_u (see* Sec. 5.1).



Then the analytic signal is

$$z_s(t) = (S_0/\pi)\int_0^{\omega_u}\exp(j\omega t)\,d\omega = (S_0/j\pi t)[\exp(j\omega_u t) - 1]$$

On splitting the above expression into its real and imaginary

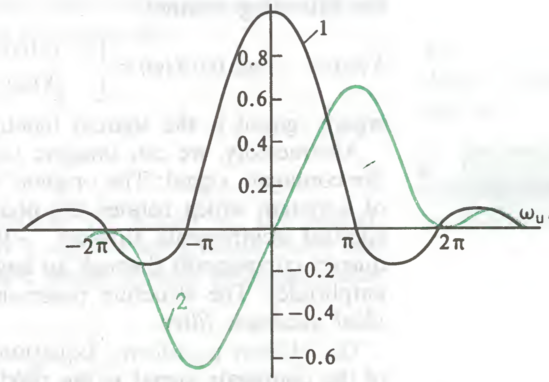


Fig. 5.4 Original and conjugate signals: (*1*) ideal baseband signal; (*2*) conjugate signal

parts, we obtain

$$s(t) = (S_0\omega_u/\pi)\frac{\sin\omega_u t}{\omega_u t}$$

(which is a previously known result), and

$$\hat{s}(t) = (S_0\omega_u/\pi)\sin^2(\omega_u t/2)/(\omega_u t/2)$$

Plots of the signals are given in Fig. 5.4.

The spectrum of the analytic signal. Let us analyse the spectrum of the analytic signal, that is, the function $Z_s(\omega)$ from which we can find $z_s(t)$ by using the inverse Fourier transform:

$$z_s(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} Z_s(\omega)\exp(j\omega t)\,d\omega$$

On the basis of Eq. (5.38), we may argue that this function is non-zero only in the region of positive frequencies:

$$Z_s(\omega) = \begin{cases} 2\,S(\omega), & \omega > 0 \\ 0, & \omega < 0 \end{cases} \tag{5.42}$$

If $\hat{S}(\omega)$ is the spectrum of the conjugate signal, then, owing to the linearity of the Fourier transform, we may write

$$Z_s(\omega) = S(\omega) + j\hat{S}(\omega) \tag{5.43}$$

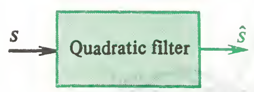
Solve Problems 13 and 14

Therefore, the equality in (5.42) will be fulfilled only when the spectra of the original and of the conjugate signals are related in

the following manner:

$$\hat{S}(\omega) = -j\,\mathrm{sgn}(\omega)\,S(\omega) = \begin{cases} -jS(\omega), & \omega > 0 \\ jS(\omega), & \omega < 0 \end{cases} \tag{5.44}$$

where sgn(ω) is the signum function.



Abstractly, we can imagine the following method for deriving the conjugate signal: The original wave $s(t)$ is applied to the input of a system which rotates the phases of all the positive-frequency spectral components through $-90°$ and of all the negative-frequency components through an angle of 90° without changing their amplitudes. The structure possessing such properties is called the *ideal quadratic filter.*

The Hilbert transform. Equation (5.44) shows that the spectrum of the conjugate signal is the product of the spectrum $S(\omega)$ of the original signal and the function $-j\,\mathrm{sgn}(\omega)$. Therefore, the conjugate signal is the convolution of two functions, $s(t)$ and $f(t)$, and this convolution is the inverse Fourier transform of $-j\,\mathrm{sgn}(\omega)$.

For mathematical convenience, let us represent the function $-j\,\mathrm{sgn}(\omega)$ as the following limit:

$$-j\,\mathrm{sgn}(\omega) = \lim_{\varepsilon \to 0}\left[-j\,\mathrm{sgn}(\omega)\exp(-\varepsilon|\omega|)\right]$$

Then,

$$f(t) = \lim_{\varepsilon \to 0}\left\{(j/2\pi)\int_{-\infty}^{0}\exp[(\varepsilon + jt)\omega]\,d\omega - (j/2\pi)\int_{0}^{\infty}\exp[(-\varepsilon + jt)\omega]\,d\omega\right\} = \frac{1}{\pi t}$$

By virtue of the foregoing, the conjugate signal is connected to the original signal by a relation of the form

The convolution of two signals is formed

$$\hat{s}(t) = s(t) * (1/\pi t) = (1/\pi)\int_{-\infty}^{\infty}\frac{s(\tau)}{t-\tau}\,d\tau \tag{5.45}$$

Alternatively, we may express the signal $s(t)$ in terms of $\hat{s}(t)$ assumed to be known. To this end, it will suffice to note that the following relation exists between the spectra

$$S(\omega) = j\,\mathrm{sgn}(\omega)\,\hat{S}(\omega)$$

Therefore, the respective equation will differ from (5.45) only in sign:

$$s(t) = -\hat{s}(t) * (1/\pi t) = \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{\hat{s}(\tau)}{\tau - t}\,d\tau \tag{5.46}$$

In mathematics, Eqs. (5.45) and (5.46) are known as the *Hilbert transform pair.* Symbolically, this is written as

Work Problems 15 and 16

$$\hat{s}(t) = H[s(t)]$$
$$s(t) = H^{-1}[\hat{s}(t)] \tag{5.47}$$

The principal value of an integral

Since the function $1/(t-\tau)$, called the kernel of the Hilbert transform, has a discontinuity at $\tau = t$, the integrals in (5.45) and (5.46) should be understood in the sense of their principal value. For example,

$$\hat{s}(t) = \frac{1}{\pi}\lim_{\xi \to 0}\left[\int_{-\infty}^{t-\xi}\frac{s(\tau)}{t-\tau}\,d\tau + \int_{t+\xi}^{\infty}\frac{s(\tau)}{t-\tau}\,d\tau\right]$$

Some properties of the Hilbert transforms. The simplest property of these integral transforms is their linearity:

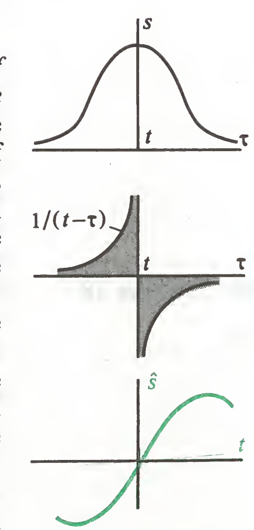
$$H[a_1 s_1(t) + a_2 s_2(t)] = a_1 H[s_1(t)] + a_2 H[s_2(t)]$$

at any values of the constants a_1 and a_2, which can be verified by inspection.

Since the kernel of the Hilbert transform is an odd function of τ about the point $\tau = t$, the signal conjugate to the constant is identically equal to zero:

$$H[\mathrm{const}] = \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{\mathrm{const}}{t-\tau}\,d\tau = 0$$

Another important property of the Hilbert transform is this: *If for any t the original signal s(t) is a maximum or a minimum, then the conjugate signal crosses zero in the vicinity of that point.* This can be verified by combining in the same drawing the plots of $s(t)$ and of the kernel $1/(t-\tau)$. Let t be close to that value of τ at which $s(t)$ is a maximum or a minimum. Then, owing to the fact that the signal shows even symmetry and the kernel shows odd symmetry the areas bounded by the horizontal axis and the curve representing the integrand function of the Hilbert transform cancel each other. Figuratively speaking, if the original signal varies in time as the cosine, the signal conjugate to it changes as the sine.



It is to be noted that the Hilbert transforms are *non-local*: the behaviour of the conjugate signal in the vicinity of any point depends on the properties of the original signal along the entire time axis, although the largest contribution comes, of course, from the vicinity of the point in question.

The Hilbert transform of a harmonic signal. For our subsequent

discussion it is important to know signals conjugate to simple harmonic waves of the form $\cos \omega t$ and $\sin \omega t$. The results can be derived by directly using Eq. (5.45). It is simpler, however, to proceed as follows. Let an arbitrary signal $s(t)$ be defined by its Fourier transform:

$$s(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} S(\omega)(\cos \omega t + j \sin \omega t)\, d\omega \tag{5.48}$$

▲ Work Problem 17

On the basis of Eq. (5.44), the Fourier transform of the conjugate signal is:

$$\hat{s}(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} -j\, \mathrm{sgn}(\omega)\, S(\omega)(\cos \omega t + j \sin \omega t)\, d\omega$$

$$= \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathrm{sgn}(\omega)\, S(\omega)(\sin \omega t - j \cos \omega t)\, d\omega \tag{5.49}$$

From a comparison of Eq. (5.49) with Eq. (5.48), the following relationships can be stated for the Hilbert transform:

$$H[\cos \omega t] = \sin \omega t\, \mathrm{sgn}(\omega)$$
$$H[\sin \omega t] = -\cos \omega t\, \mathrm{sgn}(\omega) \tag{5.50}$$

The Hilbert transform of a narrowband signal. Let a narrowband signal $s(t)$ be represented by its in-phase and quadrature amplitudes at some arbitrary reference frequency:

$$s(t) = A_s(t) \cos \omega_0 t - B_s(t) \sin \omega_0 t \tag{5.51}$$

We will analyse the properties of the signal conjugate to $s(t)$. To this end, we substitute Eq. (5.51) into Eq. (5.45), first having expanded the slowly varying functions $A_s(t)$ and $B_s(t)$ in a Taylor series about the point $\tau = t$:

▲ Solve Problem 18

$$\hat{s}(t) = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{d\tau}{t-\tau} [A_s(t) + A_s'(t)(t-\tau) + \ldots] \cos \omega_0 \tau$$

$$- \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{d\tau}{t-\tau} [B_s(t) + B_s'(t)(t-\tau) + \ldots] \sin \omega_0 \tau$$

Since the derivatives A_s', B_s', A_s'', B_s'', etc. are negligibly small, we limit ourselves to the first terms of the expansions. Then approximately

$$\hat{s}(t) = A_s(t) H[\cos \omega_0 t] - B_s(t) H[\sin \omega_0 t]$$
$$= A_s(t) \sin \omega_0 t + B_s(t) \cos \omega_0 t \tag{5.52}$$

Thus, in the case on hand the conjugate signal is likewise a narrowband wave. The last equality implies that if the complex envelope of the original signal is

$$\tilde{U}_s(t) = A_s(t) + jB_s(t)$$

then for the conjugate signal

$$\tilde{U}_{\hat{s}}(t) = B_s(t) - jA_s(t) = -j\tilde{U}_s(t)$$

Thus, the complex envelope of the conjugate signal differs from that of the original signal by a constant phase shift of 90° lagging.

The envelope, the total phase, and the instantaneous frequency. Under the Hilbert transformation, the envelope U_s of an arbitrary signal $s(t)$ is defined as a function representing time variations in the analytic signal:

Under the Hilbert transformation, the envelope and the instantaneous frequency are uniquely related to each other and may not be chosen arbitrarily

$$U_s(t) = |z_s(t)| = \sqrt{s^2(t) + \hat{s}^2(t)} \tag{5.53}$$

The validity of this definition can be checked, using the envelope of a narrowband signal as an example. Here, on the basis of Eqs. (5.51) and (5.52) the envelope is defined by

$$U_s(t) = \sqrt{A_s^2(t) + B_s^2(t)}$$

The above equation has been developed earlier in Sec. 5.3 from other considerations.

By definition, the total phase of any signal $s(t)$ is equal to the entire argument of the analytic signal $z_s(t)$:

$$\psi_s(t) = \arg z_s(t) = \arctan[\hat{s}(t)/s(t)] \tag{5.54}$$

The instantaneous frequency $\omega_s(t)$ of the signal is the time derivative of the total phase:

$$\omega_s(t) = \frac{d}{dt} \arctan \frac{\hat{s}(t)}{s(t)} = \frac{\hat{s}'(t)s(t) - s'(t)\hat{s}(t)}{\hat{s}^2 + s^2} \tag{5.55}$$

Consider several examples of how the above characteristics of various signals can be found.

Example 5.7. *Let there be a simple harmonic wave*

$s(t) = U_0 \cos \omega_0 t$

Then the conjugate signal will be

$\hat{s}(t) = U_0 \sin \omega_0 t$

The envelope of the original signal

$U_s = \sqrt{s^2(t) + \hat{s}^2(t)} = U_0$

does not, naturally, depend on time and is equal to its amplitude.

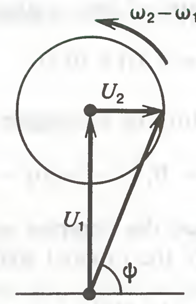
The total phase of the signal is

$\psi_s(t) = \omega_0 t$

and, finally, its instantaneous frequency is

$\omega_s = \omega_0$

This example shows that, by defining the envelope, total phase and instantaneous frequency of a signal in terms of Hilbert transforms, we obtain results which agree with our usual notions about the properties of harmonic waves.



The rate of phase change is low

Example 5.8. *The signal $s(t)$ is the sum of two harmonic waves varying in amplitude and frequency:*

$s(t) = U_1 \cos \omega_1 t + U_2 \cos \omega_2 t$

Since

$\hat{s}(t) = U_1 \sin \omega_1 t + U_2 \sin \omega_2 t$

then the envelope of the signal varies in time as

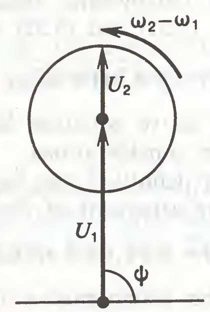
$U_s(t) = \sqrt{U_1^2 + U_2^2 + 2U_1 U_2 \cos(\omega_2 - \omega_1)t}$

The total phase of the signal is

$$\psi_s(t) = \arctan \frac{U_1 \sin \omega_1 t + U_2 \sin \omega_2 t}{U_1 \cos \omega_1 t + U_2 \cos \omega_2 t}$$

The instantaneous frequency can be found by Eq. (5.35) which yields the following result:

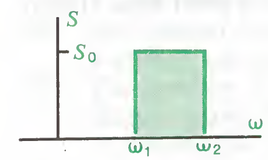
$$\omega_s(t) = \frac{\omega_1 U_1^2 + \omega_2 U_2^2 + U_1 U_2 (\omega_1 + \omega_2) \cos(\omega_2 - \omega_1) t}{U_1^2 + U_2^2 + 2U_1 U_2 \cos(\omega_2 - \omega_1) t}$$



The rate of phase change is high

The variations in the instantaneous frequency with time are due to the fact that in this case the phase of the resultant phasor representing the sum of the two harmonic waves changes at a varying rate, according as the phasors of the components are oriented relative to each other.

Example 5.9. *Consider an ideal bandpass signal band-limited to the frequency interval $[\omega_1, \omega_2]$.*



The corresponding analytic signal takes the form

$$z_s(t) = (S_0/\pi) \int_{\omega_1}^{\omega_2} \exp(j\omega t)\, d\omega = (S_0/\pi t)[(\sin \omega_2 t - \sin \omega_1 t) - j(\cos \omega_2 t - \cos \omega_1 t)]$$

The envelope of the original bandpass signal is

$$U_s(t) = (S_0/\pi t)\sqrt{(\sin \omega_2 t - \sin \omega_1 t)^2 + (\cos \omega_2 t - \cos \omega_1 t)^2} = \frac{S_0(\omega_2 - \omega_1)}{\pi} \left| \frac{\sin \frac{\omega_2 - \omega_1}{2} t}{\frac{\omega_2 - \omega_1}{2} t} \right|$$

Finally, the instantaneous frequency of the signal is

$$\omega_s = \frac{d}{dt} \left[\arctan \frac{-(\cos \omega_2 t - \cos \omega_1 t)}{\sin \omega_2 t - \sin \omega_1 t} \right]$$

After simple manipulations, we find in accordance with Eq. (5.38) that the instantaneous frequency

$\omega_s = (\omega_1 + \omega_2)/2$

is independent of time in our case and is equal to the centre frequency of the interval within which the spectrum is concentrated.

To sum up, the concept of the analytic signal enables us to define the envelope and the instantaneous frequency of narrowband signals without resort to the somewhat artificial concept of the reference frequency typical of the complex-envelope method. Moreover, Eqs. (5.53) through (5.55) remain in force when applied to arbitrary signals not necessarily quasiharmonic (narrowband). However, we cannot of course require that the envelope and the instantaneous frequency possess a simple and easy-to-grasp physical meaning.

Closing remarks. The theory of the analytic signal, as applied to the theory of oscillations and waves, was developed by Gabor [17] in the 1940s. However, the Hilbert transforms had appeared in mathematics much earlier as a technique for solving the so-called boundary problem in the theory of analytic functions [10]. The rationale of the problem is as follows.

Let $\zeta = \xi + j\eta$ be a complex variable, and $f(\zeta)$ be a complex analytic function in the upper half-plane, that is, for $\eta > 0$. On the real axis which is the boundary of the analyticity region, the function $f(\xi)$ has both a real and an imaginary part:

$f(\xi) = f_1(\xi) + jf_2(\xi)$

It is required to establish the relation that connects the functions $f_1(\xi)$ and $f_2(\xi)$. The solution to the problem is given by a Hilbert transform pair:

$$f_2(\xi) = H[f_1(\xi)]$$

$$f_1(\xi) = H^{-1}[f_2(\xi)]$$

It can be shown that the analytic signal $z_s(t)$ does possess the property of analyticity in the upper half-plane, if it is regarded as a function of a complex variable, $t = t' + jt''$. It is this property that explains the origin of the term "analytic".

In recent years, the analytic-signal methods and the Hilbert transforms have taken up a strong position as techniques of communication theory. Some of the most interesting problems in this field are covered in [24].

Summary

- Band-limited signals extend without bound in time.
- The simplest signals in this class, the ideal low-pass signal and the ideal bandpass signal, appear at the output of the corresponding ideal filters excited by a delta impulse.
- Two ideal low-pass signals can be made orthogonal through the proper choice of the time shift between them.
- The Kotelnikov series is a generalized Fourier series which yields the expansion of the signal in terms of basis functions band-limited to a certain highest frequency f_u. Here, the basis functions are ideal low-pass signals shifted in time relative to one another by an interval $1/(2f_u)$.
- The coefficients of the Kotelnikov series are the samples of the signal being expanded, taken at equal time intervals.
- If the signal spectrum has no components at frequencies higher than f_u, the Kotelnikov series gives an exact (in the mean-square sense) representation of the signal.
- The bandwidth of a narrowband signal must be a small fraction of the centre frequency. Narrowband signals are quasiharmonic; in the general case, their amplitude and frequency slowly vary in time.
- The concept of the complex envelope is a generalization of the concept of the complex amplitude to narrowband signals.
- The physical envelope is equal to the magnitude of the complex envelope. Its form does not depend on the choice of the reference frequency.
- The instantaneous frequency of a narrowband signal is the sum of the reference frequency and the time derivative of the argument of the complex envelope.
- The spectrum of a narrowband signal can be derived from that of its complex envelope by translating it by the magnitude of the reference frequency.
- To any real signal we can assign a complex analytic signal containing solely positive-frequency spectral components.

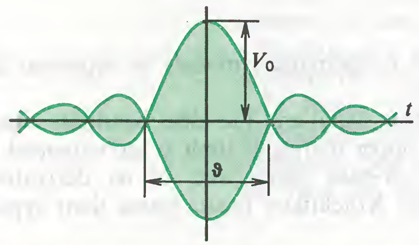
- The real part of an analytic signal is equal to the original signal. The imaginary part is called the conjugate signal.
- The relation between the original signal and its conjugate signal is established by a Hilbert transform pair.
- The envelope of an arbitrary signal is equal to the magnitude of the corresponding analytic signal. The instantaneous frequency is defined as a derivative of the argument of the analytic signal.

Review Questions

1. Why is it that band-limited signals are suitable mathematical models to represent the real waves observed in communication circuits?
2. Draw approximate waveforms of the ideal low-pass signal and the ideal bandpass signal.
3. How can it be explained that an increase in the upper frequency limit is accompanied by a rise in the extremal values of the ideal low-pass signal and of its derivative?
4. Draw plots of several functions belonging to the Kotelnikov basis. Name their typical properties.
5. Write the formula of the Kotelnikov series and state in words the Kotelnikov theorem.
6. Draw a plot for the case when samples are not taken frequently enough for the Kotelnikov series to represent a signal faithfully.
7. What controls the magnitude of the error in the signal representation by the Kotelnikov series?
8. What is the graphical meaning of the dimensionality of the band- and duration-limited signal space?
9. Draw a typical waveform for a narrowband signal.
10. Explain how the in-phase and quadrature amplitudes of a narrowband signal can be determined experimentally.
11. Formulate the properties of the physical envelope of a narrowband signal.
12. Explain how the concept of the analytic signal can be introduced.
13. How are the spectra of the original and conjugate signals related?
14. Name the basic properties of the Hilbert transform.
15. Explain how one can take the Hilbert transform of a narrowband signal.
16. Why is it that the analytic-signal method is more general than the complex-envelope method?

Problems

1. An ideal low-pass signal has a spectrum whose magnitude is 5.5×10^{-4} V s in the frequency band from zero to 25 kHz. Find the maximum instantaneous value of the signal.

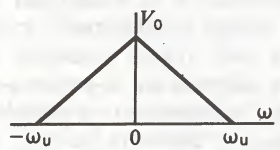


2. Measurements have shown that an ideal bandpass signal has the following parameters: $\vartheta = 20$ µs and $U_0 = 15$ V. Find the bandwidth of the signal and the magnitude of its spectrum within the bandwidth.

3. Evaluate the maximum value of the derivative of the ideal low-pass signal defined in Problem 1.

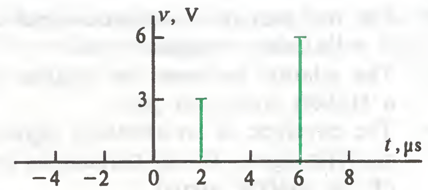
4. An automatic weather station transmits weather data at two-hour intervals. What is the highest frequency in the spectrum of the transmitted signal?

5. The spectrum $V(\omega)$ of a band-limited signal $v(t)$ is triangular in shape:



Determine the Kotelnikov coefficients for the signal, assuming that samples are taken at time intervals equal to π/ω_u.

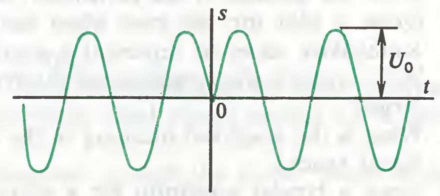
6. A band-limited signal is exactly represented by two non-zero samples:



What is the highest frequency in the spectrum of the signal? Find the instantaneous value of the signal at $t = 17$ µs.

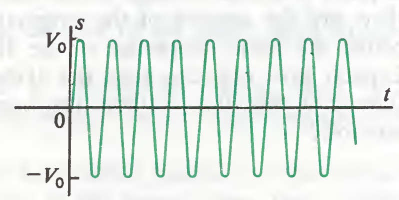
7. How will the approximation error for the signal in Example 5.3 increase, if the sampling rate is increased ten-fold?

8. A signal $s(t)$ is harmonic both at $t > 0$ and $t < 0$; at $t = 0$ its phase is reversed:



Write an expression for the complex envelope of the signal.

9. Derive an expression for the complex envelope of a truncated voltage sinewave:



Note the initial phase of the signal.

10. Determine the complex envelope of a single-tone angle-modulated signal:

$$u(t) = U_0 \cos(\omega_0 t + m \sin \Omega t)$$

11. Write an expression for the complex envelope of an LFM signal (see Chap. 4).

12. A narrowband signal in the neighbourhood of the reference frequency ω_0

has a Gaussian spectrum

$$S(\omega) = (S_0/2)\exp[-b(\omega-\omega_0)^2]$$

Determine the spectrum of the complex envelope of the signal. Find how the physical envelope changes in time. Calculate the instantaneous frequency of the signal.

Compare the results with those stated in Example 5.5. How do you explain the fundamental difference between them?

13. Find the analytic signals corresponding to simple harmonic waves $\sin \omega_0 t$ and $\cos \omega_0 t$.

14. Find the analytic signal corresponding to an r.f. pulse which has a rectangular envelope

$$u(t) = U_0[\sigma(t) - \sigma(t-\tau_p)]\cos \omega_0 t$$

15. Determine the signal conjugate to a harmonic wave $\cos \omega_0 t$ by directly using the Hilbert transform as given in (5.45).

16. Solve a problem similar to that stated in the previous case, for a signal

$$s(t) = \sin \omega_0 t/\omega_0 t$$

17. Find the signal conjugate to a periodic wave specified by its Fourier series:

$$s(t) = \sum_{k=0}^{\infty} A_k \cos(k\omega_1 t - \varphi_k)$$

18. Show that the in-phase and quadrature amplitudes of a narrowband signal $s(t)$ are connected to the components of the analytic signal by relations of the form:

$$A_s(t) = s(t)\cos \omega_0 t + \hat{s}(t)\sin \omega_0 t$$

$$B_s(t) = \hat{s}(t)\cos \omega_0 t - s(t)\sin \omega_0 t$$

Advanced Problems

19. Prove the Kotelnikov theorem in the frequency representation, which is stated thus: If a signal $s(t)$ is identically equal to zero outside the interval $t_1 < t < t_2$, then its spectrum $S(\omega)$ is uniquely defined by the sequence of values at points spaced $1/(t_2 - t_1)$ Hz apart on the frequency axis.

20. Generalize the Kotelnikov theorem to the case of band-pass signals whose spectrum occupies a frequency band $\omega_1 < \omega < \omega_2$. Find analytic expressions for the basis functions of these signals.

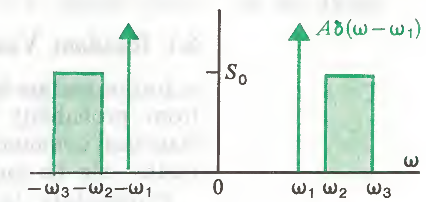
21. Let there be a narrowband signal

$$s(t) = A_s(t)\cos \omega_0 t - B_s(t)\sin \omega_0 t$$

Analyse the constraints that must be imposed on the slowly varying functions $A_s(t)$ and $B_s(t)$ in order that the instantaneous frequency be constant in time.

22. Determine and analyse the envelope, the total phase, and the instantaneous frequency of the ideal low-pass signal defined in Example 5.6.

23. Find the analytic signal for a wave, whose spectrum has a regular part and a component with a delta singularity.



24. Using the analytic-signal method, analyse the envelope and instantaneous frequency of a single-tone SSB signal.

25. Using the generalized Rayleigh formula, prove that if $s(t)$ is a finite-energy signal, the signal conjugate under the Hibert transform, $\hat{s}(t)$, is orthogonal to it.

26. Prove that the signals $s(t)$ and $\hat{s}(t)$ have the same energy and the same autocorrelation function.

27. Show that for a rectangular video pulse

$$u(t) = \begin{cases} U_0, & -\tau_p/2 < t < \tau_p/2 \\ 0, & |t| > \tau_p/2 \end{cases}$$

the conjugate signal is

$$\hat{u}(t) = \frac{1}{\pi}\ln\left|\frac{\tau_p + 2t}{\tau_p - 2t}\right|$$

Chapter 6

An Outline of the Theory of Random Signals

In radio communication systems, random signals most often occur as noise. These electromagnetic waves chaotically varying in time are produced in various physical structures when a charge carrier, say an electron, is in random motion.

In information theory, the mathematical model of a random signal is used for the probabilistic representation of the relations existing in messages having the form of meaningful texts in some language.

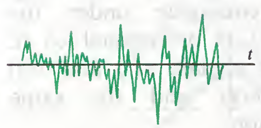
In present-day laser communication links, the signals are likewise random. The relatively high energy of a quantum of the electromagnetic field (a photon) makes it fundamentally necessary to consider the specific *quantum noise* which manifests itself in the reception of weak optical signals.

6.1 Random Variables and Their Characteristics

In this section we shall recapitulate the most important concepts from probability theory, essential in tackling the problems of statistical communication theory. A more detailed exposition of the matter can be found in [3] and [15].

Fluctuations

Probabilistic laws. A distinction of a random signal is that we cannot predict and calculate its instantaneous values in advance. From an analysis of such signals, however, we may conclude that some of its properties can be defined with sufficient accuracy in the probabilistic sense. For example, the voltage across the terminals of a noisy circuit element consists of some average level and fast time-varying random variations called *fluctuations*. These fluctuations are such that most frequently we observe relatively small variations from the average level; the larger the absolute magnitude of the variations, the more seldom they occur. Already this is a statistical pattern. If we know the probabilities of fluctuations of various magnitude, we can develop a mathematical model for the random wave involved, quite acceptable in both the theoretical and the applied sense.



Large deviations are rare

Probabilistic laws manifest themselves whenever a physical system generating a random signal is an assemblage of a very large number of smaller subsystems executing some motions more or less independent of one another. For example, the current caused to flow around a circuit by a constant-emf source owes its constancy (well known from practice) to the fact that in order to produce a current of, say, 1.6 mA, a huge number of electrons (of the order

of 10^{16}) must pass through the cross section of the conductor every second. Obviously the random fluctuations in velocity between the individual electrons can have but a negligible effect on the average current.

The electronic charge is $e = 1.6 \times 10^{-19}$ C

Probability. Present-day probability theory is an axiomatized branch of mathematics which has pooled a wealth of material amassed by science in the study of widely varying random processes.

The basis of this theory is the concept of the general population of "elementary outcomes" or random events

$$\Omega = \{A_1, A_2, ..., A_n, ...\}$$

The symbols A_i represent the likely outcomes of some random experiment. To each event $A_i \in \Omega$ is assigned a real number $P(A_i)$ which is called the *probability* of that event.

The following axioms are accepted:

The probability axioms were formulated by A. N. Kolmogorov in the 1930s

1. The probability is non-negative and does not exceed unity:

$$0 \leqslant P(A_i) \leqslant 1$$

2. The union of all events belonging to Ω is a certain event

$$\sum_{A_i \in \Omega} P(A_i) = 1$$

3. If A is some complex event, its probability is equal to the sum of all the elementary probabilities:

$$P(A) = \sum_{A_i \in A} P(A_i)$$

Measurement of probabilities. The mathematical concept of the probability of a random event is an abstract characterization associated not with the material objects of interest, but with their set-theoretic models. Some additional agreement is in order so that we can extract information about the probabilities from experimental data.

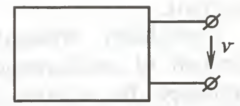
It is customary to evaluate the probability of an event in terms of the relative frequency of favourable outcomes. If we have carried out N independent trials and have observed an event A in n of them, then the *empirical* (or *sample*) *estimate* of the probability $P(A)$ that can be derived from this series of trials is

$$P_s(A) = n/N \tag{6.1}$$

Example 6.1. *The output signal from an electronic device can take on only any one of two values: $v_1 = 4.5$ V ("HIGH", event A_1) and $v_2 = 0.5$ V ("LOW", event A_2). The system can change state in a random manner at equal time intervals T. The experiment consists*

in repeatedly measuring the instantaneous value of the output signal. The instants at which the measurements are made are arbitrary, but they are spaced apart substantially more than T.

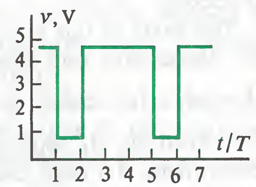
Suppose that the experimenter has carried out 100 independent trials during which the event A_1 occurred 43 times and the event A_2, 57 times. In accordance with (6.1), the sample estimates of the two probabilities are



$$P_s(A_1) = 0.43$$

and

$$P_s(A_2) = 0.57$$



From these data it does not at all follow that the theoretical (general-population) probabilities must be like that. Rather, the experimenter will advance the hypothesis that the two events are equiprobable:

▲ **Solve Problem 1**

$$P(A_1) = P(A_2) = 0.5$$

If, however, the same sample estimates are obtained in a series of 10 000 trials, the hypothesis must apparently be rejected.

The distribution function and the probability density. Let X be a random variable, that is, all possible real numbers x which take on random values in the interval $-\infty < x < +\infty$. The statistical properties of X can be described exhaustively if we have at our disposal a nonrandom function, $F(x)$, of the real argument x, which is equal to the probability that X will take any one of the allowed values x or less on any given trial of the experiment:

$$F(x) = P(X \leqslant x)$$

The function $F(x)$ is called the *distribution function* of the random variable X. If X can take on any value, then $F(x)$ is a smooth, nondiminishing function whose values lie in the interval $0 \leqslant F(x) \leqslant 1$. The following limiting equalities exist: $F(-\infty) = 0$ and $F(\infty) = 1$.

The derivative of the distribution function

$$p(x) = dF/dx$$

is the *probability density* of the random variable in question. It is obvious that

$$p(x)dx = P(x < X \leqslant x + dx)$$

that is, the term $p(x)dx$ is the probability of the random variable X falling within the interval $(x, x + dx)$.

For a continuous random variable X the probability density $p(x)$ is a smooth function. If, on the other hand, X is a discrete random variable which can take on fixed values $\{x_1, x_2, ..., x_m, ...\}$ with probabilities $\{P_1, P_2, ..., P_m, ...\}$ respectively, then

$$p(x) = \sum_i P_i \delta(x - x_i)$$

▲ **Work Problem 2**

In either case, the probability density must satisfy the condition of nonnegativity

$$p(x) \geqslant 0$$

and the condition of normality

$$\int_{-\infty}^{\infty} p(x)dx = 1$$

Averaging. Moments of a random variable. Experiments frequently yield the *averages* of a function of the random variable. If $\varphi(x)$ is a specified function of x (an outcome of a random trial), then, by definition, its average is

$$\overline{\varphi(x)} = \int_{-\infty}^{\infty} \varphi(x)p(x)dx \qquad (6.2)$$

The overscribed bar implies that the average value of the quantity under the bar is taken

From Eq. (6.2) it follows that the largest contribution to the average value comes from those values of x for which both the function being averaged, $\varphi(x)$, and the probability density $p(x)$ are simultaneously large.

Any statistical theory uses special numerical characteristics for random variables, called their *moments*. The nth-order moment of a random variable X is the nth-order average value of the random variable

● **Moments of a random variable**

$$\overline{x^n} = \int_{-\infty}^{\infty} x^n p(x)dx \qquad (6.3)$$

The first moment, which is the simplest moment of all, is the *expected value (expectation* or *mean)* of the random variable X:

$$m_1 = \bar{x} = \int_{-\infty}^{\infty} xp(x)dx \qquad (6.4)$$

Expectation generalizes the concept of the arithmetic mean in the probabilistic sense

This is an estimate for the average of a random variable, obtained from a sufficiently large series of trials.

The second moment of a random variable is the *mean square*

value of that random variable:

$$m_2 = \overline{x^2} = \int_{-\infty}^{\infty} x^2 p(x)\,dx \tag{6.5}$$

Probability theory also uses the *central moments* of random variables, which are defined as

$$\mu_n = \overline{(x-\bar{x})^n} = \int_{-\infty}^{\infty} (x-\bar{x})^n p(x)\,dx \tag{6.6}$$

Variance

The most important central moment is the *variance* of a random variable, defined by

$$\sigma_x^2 = \mu_2 = \overline{(x-\bar{x})^2} \tag{6.7}$$

It is obvious that

$$\sigma_x^2 = \overline{(x^2 - 2x\bar{x} + \bar{x}^2)} = \overline{x^2} - \bar{x}^2 \tag{6.8}$$

Standard deviation

The quantity σ_x, that is, the square root of the variance, is called the *standard deviation* of the random variable X. It is a measure for the spread of the results from individual random trials about the sample mean.

The uniform distribution is often used in error theory

The uniform distribution. Let there be a random variable X which can take on values only from the interval $x_1 \leqslant x \leqslant x_2$, such that they have an equal probability of falling within any inner subintervals of the same width Δx. Then, obviously, the probability density will be

$$p(x) = \begin{cases} 0, & x < x_1 \\ 1/(x_2 - x_1), & x_1 \leqslant x \leqslant x_2 \\ 0, & x > x_2 \end{cases}$$

The distribution function is found by integration

Solve Problems 3 and 4

$$F(x) = \int_{-\infty}^{x} p(\xi)\,d\xi = \begin{cases} 0, & x < x_1 \\ (x - x_1)/(x_2 - x_1), & x_1 \leqslant x \leqslant x_2 \\ 1, & x > x_2 \end{cases}$$

The expectation (or mean)

$$\bar{x} = \frac{1}{x_2 - x_1}\int_{x_1}^{x_2} x\,dx = \frac{(x_2 - x_1)}{2}$$

occurs, naturally, at the centre of the interval (x_1, x_2).

As can readily be verified, the variance of a random variable under the uniform distribution is

$$\sigma_x^2 = (x_2 - x_1)^2/12$$

The Gaussian (normal) distribution. In the theory of random signals, a fundamental part is played by the Gaussian probability

density

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp[-(x-m)^2/2\sigma^2] \tag{6.9}$$

defined by two numerical parameters, m and σ. The corresponding plot is a bell-shaped curve with a single maximum at the point $x = m$ (Fig. 6.1).

Note that as σ decreases, the plot localizes more and more at $x = m$

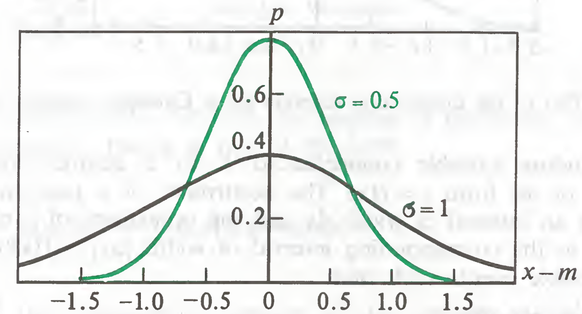


Fig. 6.1 Plot of the Gaussian probability density for several values of σ

It can be verified by direct calculation that the parameters of the Gaussian distribution have, respectively, the meaning of the expectation (mean) and the variance of a random variable:

$$\bar{x} = m$$

$$\sigma_x^2 = \sigma^2$$

The distribution function of a Gaussian random variable is

$$F(x) = (1/\sqrt{2\pi}\sigma)\int_{-\infty}^{x} \exp[-(\xi - m)^2/2\sigma^2]\,d\xi$$

By a change of variable, $t = (\xi - m)/\sigma$, it reduces to the form

$$F(x) = (1/\sqrt{2\pi})\int_{-\infty}^{(x-m)/\sigma} \exp(-t^2/2)\,dt$$

$$= \Phi[(x-m)/\sigma] \tag{6.10}$$

where the probability integral [40] is defined by

$$\Phi(x) = (1/\sqrt{2\pi})\int_{-\infty}^{x} \exp(-t^2/2)\,dt$$

The plot of the function (Fig. 6.2) is a monotonic curve varying from zero to unity.

The probability density of a function of a random variable. Let Y

Near the origin the curve has a linear portion

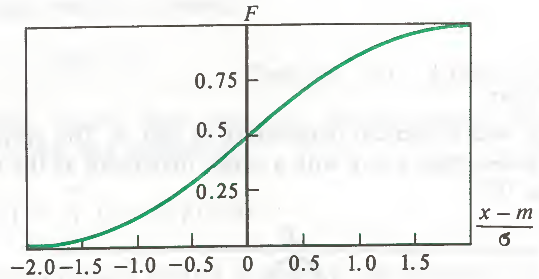


Fig. 6.2 Plot of the distribution function for a Gaussian random variable

be a random variable connected to X by a unique functional relation of the form $y=f(x)$. The occurrence of a random point x within an interval of width dx and the occurrence of a random point y in the corresponding interval of width $|dy|=|f(x)|dx$ are equiprobable events such that

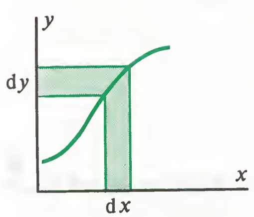
$$p_X(x)\,dx=|p_Y(y)\,dy|$$

Hence,

$$p_Y(y)=p_X(x)|dx/dy|=p_X[g(y)]\,|dg/dy| \tag{6.11}$$

where $g(y)=x$ is the inverse of $f(x)=y$.

If the functional relation between X and Y is not unique and there are several inverse functions



$$x_1=g_1(y)$$

$$x_2=g_2(y)$$

$$\cdots$$

$$x_N=g_N(y)$$

then Eq. (6.11) is generalized in the following manner:

$$p_Y(y)=\sum_{i=1}^{N}p(x_i)|dx_i/dy| \tag{6.12}$$

Example 6.2. *Linear transformation of a Gaussian random variable.* Let

$$Y=aX+b$$

and the probability density

$$p_X(x)=(1/\sqrt{2\pi}\sigma)\exp[-(x-m)^2/2\sigma^2]$$

Since

$$|dx/dy|=1/|a|$$

then, on the basis of Eq. (6.11),

$$p_Y(y)=(1/\sqrt{2\pi}\sigma|a|)\exp[-(y-b-ma)^2/2a^2\sigma^2]$$

Thus, the random variable retains its Gaussian behaviour under linear transformation. The variable derived by this transformation has the mean

$$\bar{y}=b+ma$$

and the variance

$$\sigma_y^2=a^2\sigma_x^2$$

▲ **Solve Problem 7**

The characteristic function. An average of particular importance to probability theory is one of the form

$$\Theta(v)=\overline{\exp(jvx)}=\int_{-\infty}^{\infty}p(x)\exp(jvx)\,dx \tag{6.13}$$

called the *characteristic function* of the random variable X. To within the constant coefficient, the function $\Theta(v)$ is the Fourier transform of the probability density, so

$$p(x)=\frac{1}{2\pi}\int_{-\infty}^{\infty}\Theta(v)\exp(-jvx)\,dv \tag{6.14}$$

It is clear that one and the same random variable can be equally well described by its characteristic function and by its probability density function. The choice is a matter of mathematical convenience.

Omitting elementary manipulations, here are some results:

– for a random variable with a uniform distribution over the interval $0\leqslant x\leqslant a$,

$$\Theta(v)=[\exp(jav)-1]/jav \tag{6.15}$$

– for a Gaussian random variable with specified parameters m and σ,

$$\Theta(v)=\exp(jmv-\sigma^2v^2/2) \tag{6.16}$$

The characteristic function offers a simple means for determining the moments of random variables. To demonstrate,

$$d^n\Theta(v)/dv^n=j^n\int_{-\infty}^{\infty}x^n p(x)\exp(jvx)\,dx$$

On setting $v=0$ and equating to (6.3), we get

$$m_n=j^{-n}\Theta^{(n)}(0) \tag{6.17}$$

Near the origin the curve has a linear portion

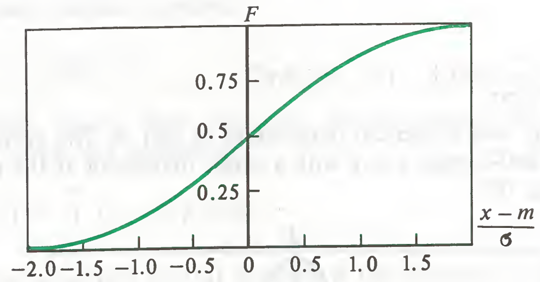


Fig. 6.2 Plot of the distribution function for a Gaussian random variable

be a random variable connected to X by a unique functional relation of the form $y = f(x)$. The occurrence of a random point x within an interval of width $\mathrm{d}x$ and the occurrence of a random point y in the corresponding interval of width $|\mathrm{d}y| = |f(x)|\,\mathrm{d}x$ are equiprobable events such that

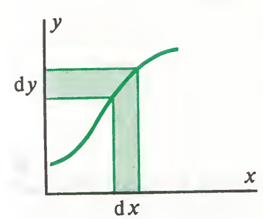
$$p_X(x)\mathrm{d}x = |p_Y(y)\mathrm{d}y|$$

Hence,

$$p_Y(y) = p_X(x)|\mathrm{d}x/\mathrm{d}y| = p_X[g(y)]\,|\mathrm{d}g/\mathrm{d}y| \tag{6.11}$$

where $g(y) = x$ is the inverse of $f(x) = y$.

If the functional relation between X and Y is not unique and there are several inverse functions



$$x_1 = g_1(y)$$

$$x_2 = g_2(y)$$

$$\cdot\quad\cdot\quad\cdot\quad\cdot$$

$$x_N = g_N(y)$$

then Eq. (6.11) is generalized in the following manner:

$$p_Y(y) = \sum_{i=1}^{N} p(x_i)\,|\mathrm{d}x_i/\mathrm{d}y| \tag{6.12}$$

Example 6.2. *Linear transformation of a Gaussian random variable.* Let

$$Y = aX + b$$

and the probability density

$$p_X(x) = (1/\sqrt{2\pi}\sigma)\exp[-(x-m)^2/2\sigma^2]$$

Since

$$|\mathrm{d}x/\mathrm{d}y| = 1/|a|$$

then, on the basis of Eq. (6.11),

$$p_Y(y) = (1/\sqrt{2\pi}\sigma|a|)\exp[-(y-b-ma)^2/2a^2\sigma^2]$$

Thus, the random variable retains its Gaussian behaviour under linear transformation. The variable derived by this transformation has the mean

$$\bar{y} = b + ma$$

and the variance

$$\sigma_y^2 = a^2\sigma_x^2$$

Solve Problem 7

The characteristic function. An average of particular importance to probability theory is one of the form

$$\Theta(v) = \overline{\exp(jvx)} = \int_{-\infty}^{\infty} p(x)\exp(jvx)\,\mathrm{d}x \tag{6.13}$$

called the *characteristic function* of the random variable X. To within the constant coefficient, the function $\Theta(v)$ is the Fourier transform of the probability density, so

$$p(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \Theta(v)\exp(-jvx)\,\mathrm{d}v \tag{6.14}$$

It is clear that one and the same random variable can be equally well described by its characteristic function and by its probability density function. The choice is a matter of mathematical convenience.

Omitting elementary manipulations, here are some results:

– for a random variable with a uniform distribution over the interval $0 \leqslant x \leqslant a$,

$$\Theta(v) = [\exp(jav) - 1]/jav \tag{6.15}$$

– for a Gaussian random variable with specified parameters m and σ,

$$\Theta(v) = \exp(jmv - \sigma^2 v^2/2) \tag{6.16}$$

The characteristic function offers a simple means for determining the moments of random variables. To demonstrate,

$$\mathrm{d}^n\Theta(v)/\mathrm{d}v^n = j^n\int_{-\infty}^{\infty} x^n p(x)\exp(jvx)\,\mathrm{d}x$$

On setting $v = 0$ and equating to (6.3), we get

$$m_n = j^{-n}\Theta^{(n)}(0) \tag{6.17}$$

Using the characteristic functions, it is convenient to find the probability densities of random variables that have been subjected to functional transformations. Thus, if

$$y = f(x)$$

then

$$\Theta_y(v) = \overline{\exp(jvy)} = \overline{\exp[jvf(x)]}$$

Work Problem 6

The problem on hand will be solved, if we are able to evaluate the Fourier transform in (6.14).

Example 6.3. *Let* $y = U_0 \cos x$, *where* U_0 *is constant, whereas* x *is the value of a random variable uniformly distributed over the interval* $-\pi \leqslant x \leqslant \pi$.

Since

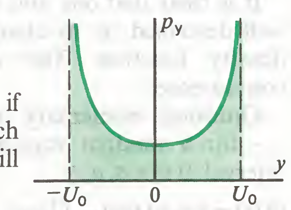
$$p_x(x) = 1/2\pi$$

then

$$\Theta_y(v) = (1/2\pi) \int_{-\pi}^{\pi} \exp(jvU_0 \cos x)\,dx = J_0(vU_0)$$

where J_0 is the Bessel function of first kind, of order zero. Using the tabulated integral [40], we get

$$p_y(y) = \frac{1}{2\pi} \int_{-\infty}^{\infty} J_0(vU_0)\exp(jvy)\,dv = \begin{cases} \dfrac{1}{\pi\sqrt{U_0^2 - y^2}}, & |y| \leqslant U_0 \\ 0, & |y| > U_0 \end{cases}$$



From the shape of the probability density plot it is seen that if we carry out a large series of trials, with the value of x taken each time at random from the specified interval, the term $U_0 \cos x$ will most often assume values close to $\pm U_0$.

6.2 Statistical Characteristics of Two and More Random Variables

As is customary, the properties of a random signal process are analysed from the collection of values observed at different fixed instants, rather than at one particular time. What follows is a brief outline of the theory of such *multivariate* or *vector systems*.

The distribution function and the probability density. Let there be random variables $\{X_1, X_2, ..., X_n\}$ which form an n-dimensional (or vector) random variable $\bar{X}$. By analogy with the one-dimensional case, the distribution function of this vector variable may be defined as

$$F(x_1, x_2, ..., x_n) = P(X_1 \leqslant x_1, X_2 \leqslant x_2, ..., X_n \leqslant x_n)$$

The corresponding n-dimensional (or nth-order multivariate) probability density $p(x_1, x_2, ..., x_n)$ satisfies the relation

$$p(x_1, x_2, ..., x_n)\,dx_1 dx_2 ... dx_n = P\{x_1 < X_1 \leqslant x_1 + dx_1, ..., x_n < X_n \leqslant x_n + dx_n\}$$

Obviously, the distribution function can be evaluated by integrating the probability density function:

$$F(x_1, x_2, ..., x_n) = \int_{-\infty}^{x_1} ... \int_{-\infty}^{x_n} p(\xi_1, ..., \xi_n)\,d\xi_1 ... d\xi_n$$

Any multivariate density has the usual properties of a univariate probability density

$$p(\xi_1, ..., \xi_n) \geqslant 0; \quad \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} p\,d\xi_1 ... d\xi_n = 1$$

If we know the n-dimensional density, we can always find the m-dimensional density for $m < n$ by integrating over the "redundant" coordinates:

$$p(\xi_1, ..., \xi_m) = \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} p(\xi_1, ..., \xi_m, ..., \xi_n)\,d\xi_{m+1} ... d\xi_n$$

Moments. If we know the corresponding multivariate probability density, we can find the various averages of any combinations of the random variables involved, and, notably, their moments. For example, limiting ourselves to the two-dimensional case as most important for our further discussion, we find by analogy with (6.4) and (6.7) the expectations (or means)

$$\bar{x}_1 = \iint_{-\infty}^{\infty} x_1 p(x_1, x_2)\,dx_1 dx_2$$
$$\bar{x}_2 = \iint_{-\infty}^{\infty} x_2 p(x_1, x_2)\,dx_1 dx_2 \quad (6.18)$$

and the variances:

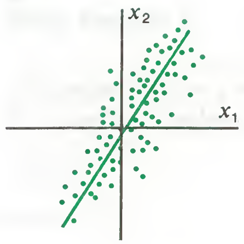
$$\sigma_1^2 = \iint_{-\infty}^{\infty} (x_1 - \bar{x}_1)^2 p(x_1, x_2)\,dx_1 dx_2$$
$$\sigma_2^2 = \iint_{-\infty}^{\infty} (x_2 - \bar{x}_2)^2 p(x_1, x_2)\,dx_1 dx_2 \quad (6.19)$$

The covariance moment

What is new in comparison with the one-dimensional case is that we can form a second-order joint moment called the *covariance moment** of a two-dimensional random variable:

$$k_{12} = \overline{x_1 x_2} = \iint_{-\infty}^{\infty} x_1 x_2 p(x_1, x_2)\, dx_1 dx_2 \tag{6.20}$$

Correlation. Suppose that we have carried out a series of trials each of which has yielded a two-dimensional random variable $\{x_1, x_2\}$. Let us agree to represent the outcome of each trial by a point on a Cartesian plane.



It may so happen that on the average the representative points fall on a straight line, which means that on any one trial the values x_1 and x_2 have the same sign. This suggests the existence of a statistical association between x_1 and x_2, called *correlation.*

On the other hand, it may so happen that the points lie in a chaotic fashion on the plane. Now the values are said to be *uncorrelated*, that is, there is no consistent statistical association between them.

The correlation moment

Quantitatively, the degree of statistical association between two random variables is stated in terms of their covariance moment k_{12} or, which is often more convenient, in terms of their *correlation moment* K_{12}** defined as the mean of the product $(x_1 - \bar{x}_1)(x_2 - \bar{x}_2)$:

$$K_{12} = \iint_{-\infty}^{\infty} (x_1 - \bar{x}_1)(x_2 - \bar{x}_2) p(x_1, x_2)\, dx_1 dx_2$$
$$= k_{12} - \bar{x}_1 \bar{x}_2 \tag{6.21}$$

We also introduce a dimensionless *correlation coefficient*

$$R_{12} = K_{12}/\sigma_1 \sigma_2 \tag{6.22}$$

When the two random variables are the same, that is, $x_1 = x_2$,

$$K_{11} = K_{22} = \sigma^2 \quad \text{and} \quad R_{11} = R_{22} = 1$$

If a random vector variable has a dimension of more than two, we may form the respective cross-correlation moments

$$K_{ij} = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} (x_i - \bar{x}_i)(x_j - \bar{x}_j) p(x_1, \dots, x_n)\, dx_1 \dots dx_n$$

$i, j = 1, 2, \dots, n,$

* Other authors [16] call it *correlation.– Translator's note.*
** This is frequently termed *covariance.* See [16, 20].– *Translator's note.*

and the respective correlation coefficients

$R_{ij} = K_{ij}/\sigma_i \sigma_j$

which can be combined into the corresponding matrices

$$\underline{K} = \begin{bmatrix} K_{11} & \dots & K_{1n} \\ K_{21} & \dots & K_{2n} \\ \dots & \dots & \dots \\ K_{n1} & \dots & K_{nn} \end{bmatrix} \quad \text{and} \quad \underline{R} = \begin{bmatrix} 1 & R_{12} \dots & R_{1n} \\ R_{21} & 1 \dots & R_{2n} \\ \dots & \dots & \dots \\ R_{n1} & \dots & 1 \end{bmatrix}$$

It can be shown that $|R_{ij}| \leqslant 1$ always, and the equality holds only when $x_1 = \pm x_j$ (completely correlated variables).

Statistical independence of random variables. By definition, the random variables $X_1, X_2, \dots, X_n$ are *statistically independent* if their multivariate probability density can be represented as the product of the corresponding univariate densities:

Statistical independence

$$p(x_1, x_2, \dots, x_n) = p(x_1) p(x_2) \dots p(x_n) \tag{6.23}$$

Statistically independent random variables are pairwise uncorrelated. To demonstrate, for $i \neq j$,

$$K_{ij} = \int_{-\infty}^{\infty} (x_i - \bar{x}_i) p(x_i)\, dx_i \int_{-\infty}^{\infty} (x_j - \bar{x}_j) p(x_j)\, dx_j = 0$$

Work Problem 12

The converse is not true in general: If random variables are uncorrelated, this does not mean automatically that they are statistically independent.

Functional transformations of multidimensional random variables. Let two random vector variables $\vec{X}$ and $\vec{Y}$ be related uniquely by the following relation:

$$y_1 = f_1(x_1, \dots, x_n)$$
$$\dots\dots$$
$$y_n = f_n(x_1, \dots, x_n)$$

for which we know the inverse functions

$$x_1 = g_1(y_1, \dots, y_n)$$
$$\dots\dots$$
$$x_n = g_n(y_1, \dots, y_n)$$

The original probability density $p_{\text{orig}}(x_1, \dots, x_n)$ is specified in advance. So that we could generalize Eq. (6.11) to a multidimensional case and evaluate the probability density of the

transformed vector variable, $p_{\text{trans}}(y_1, ..., y_n)$, we should use the Jacobian of the transformation:

The Jacobian is a coefficient of proportionality between elementary volumes in a functional transformation

$$D = \begin{vmatrix} \dfrac{\partial g_1}{\partial y_1} & \dfrac{\partial g_1}{\partial y_2} & \cdots & \dfrac{\partial g_1}{\partial y_n} \\ \cdot & \cdot & \cdots & \cdot \\ \dfrac{\partial g_n}{\partial y_1} & \dfrac{\partial g_n}{\partial y_2} & \cdots & \dfrac{\partial g_n}{\partial y_n} \end{vmatrix} \tag{6.24}$$

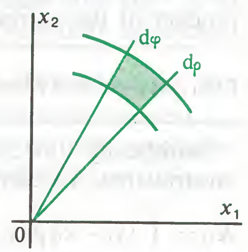
Then the sought probability function takes the form

$$p_{\text{trans}}(y_1, ..., y_n) = p_{\text{orig}}(g_1, ..., g_n)\,|D| \tag{6.25}$$

Example 6.4. *Let x_1 and x_2 be the random coordinates of the vector tip on a plane.*

If we change to the polar coordinates (ρ, φ),

$$x_1 = \rho\cos\varphi \quad \begin{cases} 0 \leqslant \rho < \infty \\ 0 \leqslant \varphi < 2\pi \end{cases}$$
$$x_2 = \rho\sin\varphi$$



then the Jacobian of the above transformation will be

$$D = \begin{vmatrix} \cos\varphi & -\rho\sin\varphi \\ \sin\varphi & \rho\cos\varphi \end{vmatrix} = \rho$$

So, if the specified probability density is $p_{\text{orig}}(x_1, x_2)$, then

$$p_{\text{trans}}(\rho, \varphi) = \rho\, p_{\text{orig}}(\rho\cos\varphi,\ \rho\sin\varphi)$$

The multivariate Gaussian distribution. Suppose that for an n-dimensional (vector) random variable, $\vec{X} = \{X_1, ..., X_n\}$, we know the means $(m_1, ..., m_n)$, the variances $(\sigma_1^2, ..., \sigma_n^2)$, and the correlation coefficient matrix $\underline{R}$.

In the general case, this information is not a sufficient basis on which to form the n-dimensional probability density. The only exception is the case where $\vec{X}$ is a multidimensional Gaussian variable. Then, by definition,

$$p(x_1, ..., x_n) = \frac{1}{\sigma_1 \ldots \sigma_n (2\pi)^{n/2} D^{1/2}} \times \exp\left(-\frac{1}{2D}\sum_{i,j=1}^{n} D_{ij}\frac{x_i - m_i}{\sigma_i}\,\frac{x_j - m_j}{\sigma_j}\right) \tag{6.26}$$

where D is the determinant of the matrix $\underline{R}$, and D_{ij} is the cofactor of the element R_{ij} in that matrix.

Let the vector variable $\vec{X}$ be formed by uncorrelated random variables such that in the matrix $\underline{R}$ only the elements on the principal diagonal are non-zero: $R_{ij} = \delta_{ij}$. Then $D = 1$ and $D_{ij} = \delta_{ij}$. On substituting them in (6.26), we have

The Kronecker delta

$$\delta_{ij} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$$

$$p(x_1, ..., x_n) = \frac{1}{(2\pi)^{n/2}\sigma_1 \ldots \sigma_n}\exp\left[-\frac{1}{2}\sum_{i=1}^{n}\frac{(x_i - m_i)^2}{\sigma_i^2}\right] = p(x_1)\,p(x_2)\ldots p(x_n)$$

where each of the univariate Gaussian distributions has the mean m_i and the standard deviation σ_i. Hence we may state an important property of the Gaussian distribution: *If a Gaussian set is formed by uncorrelated random variables, they are all statistically independent.*

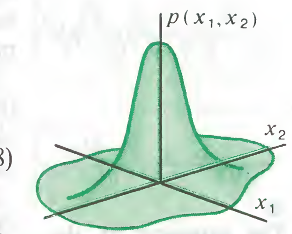
▲ Solve Problem 10

In our further discussion we shall frequently use the bivariate Gaussian probability density

$$p(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1 - R^2}}\exp\left\{-\frac{1}{2(1 - R^2)}\left[\frac{(x_1 - m_1)^2}{\sigma_1^2} - 2R\frac{(x_1 - m_1)(x_2 - m_2)}{\sigma_1\sigma_2} + \frac{(x_2 - m_2)^2}{\sigma_2^2}\right]\right\} \tag{6.27}$$

where $R = R_{12} = R_{21}$ is the correlation coefficient between x_1 and x_2. The above expression is simplified if $m_1 = m_2 = 0$ and $\sigma_1 = \sigma_2 = \sigma$:

$$p(x_1, x_2) = \frac{1}{2\pi\sigma^2\sqrt{1 - R^2}}\exp\left[-\frac{1}{2(1 - R^2)\sigma^2} \times (x_1^2 - 2Rx_1x_2 + x_2^2)\right] \tag{6.28}$$



This probability density is depicted by a smooth surface constructed over the coordinate plane (x_1, x_2). The value of $p(x_1, x_2)$ is an absolute maximum at the origin. The configuration of the surface is decided by the correlation coefficient R.

The multidimensional characteristic function. A generalization of the characteristic function to the multidimensional case is the n-dimensional Fourier transform of the corresponding probability density:

$$\Theta(v_1, v_2, ..., v_n) = \overline{\exp\left[j(x_1v_1 + x_2v_2 + ... + x_nv_n)\right]} = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\exp\left[j(x_1v_1 + ... + x_nv_n)\right] \times p(x_1, \ldots, x_n)\,dx_1 \ldots dx_n \tag{6.29}$$

The multidimensional characteristic function defines a system of random variables to the same degree of completeness as does the inverse Fourier transform of the corresponding probability density:

$$p(x_1, ..., x_n) = \frac{1}{(2\pi)^n} \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} \Theta(v_1, ..., v_n) \times \exp[-j(x_1 v_1 + ... + x_n v_n)]\, dv_1 ... dv_n \qquad (6.30)$$

If $\{X_1, ..., X_n\}$ are statistically independent, then, by virtue of Eq. (6.29), the multidimensional characteristic function factors into the one-dimensional characteristic functions associated with the individual random variables:

$$\Theta(v_1, ..., v_n) = \prod_{i=1}^{n} \Theta_i(v_i) \qquad (6.31)$$

It can be shown that the characteristic function of a multidimensional Gaussian random variable $\vec{X} = \{X_1, ..., X_n\}$ has the form

$$\Theta(v_1, ..., v_n) = \exp\left(j \sum_{k=1}^{n} m_k v_k - \frac{1}{2} \sum_{k,l=1}^{n} \sigma_k \sigma_l R_{kl} v_k v_l\right) \qquad (6.32)$$

where m_k and σ_k^2 are the mean and variance of the random vector variable $\vec{X}_k$, and R_{kl} is an element of the correlation coefficient matrix.

The probability density of the sum of random variables. If in (6.29) we set $v_1 = v_2 = ... = v_n = v$ then the multidimensional characteristic function changes into the one-dimensional characteristic function of the sum $x_1 + x_2 + ... + x_n$:

The property of the characteristic function

$$\Theta_\Sigma(v) = \exp[jv(x_1 + x_2 + ... + x_n)]$$

Hence, on taking the inverse Fourier transform, we can find the probability density of the sum. For example, if $\{X_1, ..., X_n\}$ are Gaussian, uncorrelated (and, in consequence, statistically independent) random variables, each of mean value m_k and of standard deviation σ_k, then it follows from (6.32) that

$$\Theta_\Sigma(v) = \exp\left(jv \sum_{k=1}^{n} m_k - \frac{1}{2} v^2 \sum_{k=1}^{n} \sigma_k^2\right) \qquad (6.33)$$

By comparing Eq. (6.33) with Eq. (6.16), we can see that the sum of Gaussian random variables is likewise normally distributed, whereas the means and the variances are respectively added

together:

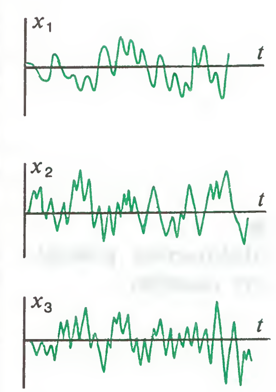
$$m_\Sigma = \sum_{k=1}^{n} m_k, \qquad \sigma_\Sigma^2 = \sum_{k=1}^{n} \sigma_k^2 \qquad (6.34)$$

The central limit theorem

In probability theory, a far stronger assertion is proved, which is known as Lyapunov's *Central Limit Theorem* [15]. This theorem states that subject to certain constraints usually satisfied in physical systems, the distribution of the sum of n independent random variables whose variances are finite and whose probability distributions are arbitrary tends to the Gaussian distribution as $n \to \infty$.

6.3 Random Processes

The theory of random variables deals with probabilistic phenomena "in statics", that is, by treating them as some fixed experimental results. However, the techniques of the classical probability theory prove inadequate when it comes to signals representing random phenomena varying in time. They are tackled by a special branch of mathematics, called the *theory of random processes*.



By definition, a *random process* $X(t)$ is a function characterized by the fact that at any time t the values it assumes are random variables.

Ensembles of realizations. When we deal with deterministic signals, we represent them by functional relations or by waveforms. The situation is far more complicated when it comes to random processes. By noting the instantaneous values of a random process within a certain time interval, we obtain only a single *realization* of the random process. Theoretically, a random process is expressed in terms of an infinite set of such realizations which form a *statistical ensemble*. An example of such an ensemble is the set of signals $\{x_1(t), x_2(t), ..., x_k(t), ...\}$ which can simultaneously be observed at the outputs of absolutely indentical noise voltage generators.

An ensemble of realizations

It is not mandatory for the realizations of a random process to be represented by functions with a complicated, irregular time behaviour. Frequently, we have to do with random processes formed by, say, all kinds of harmonic signals, $U \cos(\omega t + \varphi)$, for which one of the parameters, U, ω or φ, is a random variable assuming a particular value in each realization. The random character of such a signal is due to the fact that we cannot predict the value of that parameter in advance, prior to experiment.

Quasideterministic random processes

Random processes whose realizations depend on a finite number of parameters are usually termed *quasideterministic*.

The probability densities of random processes. Let $X(t)$ be

a random process defined by an ensemble of realizations, and t_1 be an arbitrary point of time. By noting the values $\{x_1(t_1), x_2(t_1), \ldots, x_k(t_1)\}$, assumed by the individual realizations, we take a *one-dimensional section* through the random process in question and observe the random variable $X(t_1)$. Its probability density $p(x, t_1)$ is called the *one-dimensional* or *univariate probability density* of the process $X(t)$ at time t_1. By definition, $dP = p(x, t_1)dx$ is the probability that the realizations of the random process will take at time t_1 the values lying in the interval $(x, x + dx)$.

One-dimensional (univariate) probability density

The information that can be extracted from a one-dimensional (univariate) probability density is insufficient for us to judge how the realization of the random process changes with time. A far more complete description can be obtained by taking two sections through the random process at different instants of time, t_1 and t_2. The two-dimensional random variable $\{X(t_1), X(t_2)\}$ emerging from this mental experiment can be represented by a *two-dimensional* or *bivariate probability density*, $p(x_1, x_2, t_1, t_2)$. Using this characteristic, we can determine the probability of the event that the realization of the random process at $t = t_1$ occurs in the neighbourhood of point $x = x_1$, and at $t = t_2$, in the neighbourhood of point $x = x_2$.

A natural generalization is the *n-dimensional* (*nth-order multivariate*) *section* through the random process ($n > 2$), resulting in the *n-dimensional* or *nth-order multivariate probability density*, $p(x_1, x_2, \ldots, x_n, t_1, t_2, \ldots, t_n)$.

Multivariate probability densities

The multivariate probability density of a random process must satisfy the usual conditions imposed on the probability density of a set of random variables (see Sec. 6.2). Also, the quantity $p(x_1, x_2, \ldots, x_n, t_1, t_2, \ldots, t_n)$ must be independent of the order in which its arguments are arranged (the condition of symmetry).

Sometimes, instead of the n-dimensional probability density, use is made of the n-dimensional characteristic function which is related to the correspondent probability density by the Fourier transform:

$$\Theta(v_1, v_2, \ldots, v_n, t_1, t_2, \ldots, t_n)$$
$$= \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} p(x_1, \ldots, x_n, t_1, \ldots, t_n) \exp[j(v_1 x_1 + \ldots$$
$$+ v_n x_n)] dx_1 \ldots dx_n \tag{6.35}$$

Multivariate probability densities of a sufficiently high order make it possible to describe the properties of random processes to a sufficient degree of detail. However, such densities are frequently very difficult to obtain and analyse.

Moment functions of random processes. Less detailed, but satisfactory characteristics (in the practical sense) of random processes can be obtained by determining the moments of those random variables that are encountered in sections through the processes. Since in the general case these moments are functions of time arguments, they are called *moment functions*.

In statistical communication theory, three moment functions of the lowest order have come to play the most important role. They are the *expectation* (or *mean*), the *variance*, and the *autocorrelation function*.

The expectation

$$m(t) = \overline{x(t)} = \int_{-\infty}^{\infty} x p(x, t) dx \tag{6.36}$$

is the mean value of the process $X(t)$ at the present instant of time t; averaging is done over the entire ensemble of process realizations.

The variance

$$\sigma^2(t) = \overline{[x(t) - m(t)]^2} = \int_{-\infty}^{\infty} [x(t) - m(t)]^2 p(x, t) dx \tag{6.37}$$

is a measure of the spread of the instantaneous values taken on by individual realizations in a fixed section about their mean.

The second central moment

$$K(t_1, t_2) = \overline{[x(t_1) - m(t_1)][x(t_2) - m(t_2)]}$$
$$= \int \int_{-\infty}^{\infty} [x(t_1) - m(t_1)][x(t_2) - m(t_2)] p(x_1, x_2, t_1, t_2) dx_1 dx_2 \tag{6.38}$$

is called the autocorrelation function* of the random process $X(t)$. This function gives information about how the values of $X(t)$ at one time, t_1, are related to those at the other time, t_2, in the statistical sense.

The autocorrelation function

From a comparison of (6.37) and (6.38), it will be noted that when the time sections coincide, the autocorrelation function is numerically equal to the variance of the random process:

$$K(t_1, t_2)|_{t_1 = t_2 = t} = \sigma^2(t) \tag{6.39}$$

Stationary random processes. The term *stationary* refers to the fact that the statistical characteristics of a process remain unchanged with time. Random signals which are typical realizations of stationary random processes form an important and widely encountered class of random oscillations.

* Some authors call it the *autocovariance function*. See [20].– *Translator's note.*

A random process is said to be *narrow-sense stationary* if any of its nth-order multivariate probability densities is invariant under a time shift τ:

$$p(x_1, \ldots, x_n, t_1, \ldots, t_n) = p(x_1, \ldots, x_n, t_1 + \tau, \ldots, t_n + \tau) \tag{6.40}$$

If we limit ourselves to the requirement that the mean m and the variance σ^2 be independent of time, but let the autocorrelation function be only dependent on the time difference $\tau = |t_2 - t_1|$, that is

$$K(t_1, t_2) = K(\tau)$$

Wide-sense and narrow-sense stationarity

then the random process will be *wide-sense stationary.* It is obvious that the property of wide-sense stationarity stems from the narrow-sense stationarity, but not the other way around.

As follows from the definition of a stationary random process, the autocorrelation function shows even symmetry:

$$K(\tau) = K(-\tau)$$

Also, the values of this function at any τ do not exceed the values at $\tau = 0$:

$$K(\tau) \leqslant K(0) = \sigma^2 \tag{6.41}$$

The proof is as follows. From the obvious inequality

$$\overline{\{[x(t) - m] - [x(t+\tau) - m]\}^2} \geqslant 0$$

it follows that

$$\overline{[x(t) - m]^2} - 2\overline{[x(t) - m][x(t+\tau) - m]} + \overline{[x(t+\tau) - m]^2}$$
$$= 2\sigma^2 - 2K(\tau) \geqslant 0$$

from which Eq. (6.41) results immediately.

The correlation coefficient of a random process

Frequently it is convenient to introduce the normalized autocorrelation function

$$R(\tau) = K(\tau)/\sigma^2 \tag{6.42}$$

also called the *correlation coefficient* of a stationary random process, for which $R(0) = 1$.

To illustrate the concept of a stationary random process, let us consider two examples.

Example 6.5. *A random process $U(t)$ is formed by realizations of the form $u(t) = U_0 \cos(\omega_0 t + \varphi)$, where U_0 and ω_0 are constant, whereas the phase angle φ is a random variable uniformly distributed over the interval $-\pi \leqslant \varphi \leqslant \pi$.*

Since the probability density function of the phase angle is

$$p_\varphi = 1/2\pi$$

the expectation (mean) of the process will be

$$\bar{u} = \overline{U_0 \cos(\omega_0 t + \varphi)} = (U_0/2\pi) \int_{-\pi}^{\pi} \cos(\omega_0 t + \varphi)\, d\varphi = 0$$

Similarly, the variance is

$$\sigma^2 = \overline{(u - \bar{u})^2} = U_0^2 \overline{\cos^2(-\omega_0 t + \varphi)} = U_0^2/2$$

Finally, the autocorrelation function is

$$K(t_1, t_2) = U_0^2 \overline{\cos(\omega_0 t_1 + \varphi)\cos(\omega_0 t_2 + \varphi)}$$
$$= (U_0^2/2)\{\overline{\cos[\omega_0(t_1 + t_2) + 2\varphi]} + \cos\omega_0(t_2 - t_1)\}$$
$$= (U_0^2/2)\cos\omega(t_2 - t_1)$$

Solve Problem 9

Thus, the random process in question satisfies all the conditions necessary to assure wide-sense stationarity.

Example 6.6. *Consider a random process made up of realizations $u(t) = U_0 \cos(\omega_0 t + \varphi_0)$, such that ω_0 and φ_0 are fixed numbers, and U_0 is a random variable distributed in an arbitrary fashion.*

The expectation

$$\bar{u} = \bar{U}_0 \cos(\omega_0 t + \varphi_0)$$

will be independent of time only when $\bar{U}_0 = 0$. Therefore, in the general case the random process in question will be nonstationary.

Most random processes in telecommunications are ergodic

Ergodicity. A stationary random process $X(t)$ is called *ergodic*, if its ensemble averages may be replaced with its time averages. Averaging is performed over a single realization $x(t)$ whose duration T tends to infinity. Denoting a time average by angle quotes, $\langle\ \rangle$, we will write the expectation (mean) of an ergodic random process as

$$m = \langle x(t)\rangle = \lim_{T\to\infty} \frac{1}{T}\int_0^T x(t)\, dt \tag{6.43}$$

which is equal to the constant term of the selected realization. The variance of such a process is

$$\sigma^2 = \langle[x(t) - m]^2\rangle = \lim_{T\to\infty} \frac{1}{T}\int_0^T [x(t) - m]^2\, dt = \langle x^2(t)\rangle - m^2 \tag{6.44}$$

The physical significance of the variance of a random process

Since $\langle x^2\rangle$ is the mean power in a realization, and m^2 is the power in the constant term, *the variance has the easy-to-grasp*

meaning of the power in the fluctuating component of the ergodic process.

The autocorrelation function is found in a similar way:

$$K(\tau) = \langle [x(t) - m][x(t+\tau) - m] \rangle = \langle x(t)x(t+\tau) \rangle - m^2$$

$$= \lim_{T \to \infty} \frac{1}{T} \int_0^T x(t)x(t+\tau)\,dt - m^2 \qquad (6.45)$$

For a random process to be ergodic, it must above all be wide-sense stationary. A sufficient condition for ergodicity is that the autocorrelation function should tend to zero as the time translation τ increases without bound:

$$\lim_{\tau \to \infty} K(\tau) = 0 \qquad (6.46)$$

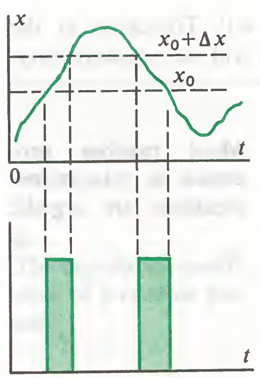
It is proved in mathematics that the above requirement can be relaxed. As has been found, a random process is ergodic, if the Slutsky condition [14] is satisfied:

$$\lim_{T \to \infty} \frac{1}{T} \int_0^T K(\tau)\,d\tau = 0 \qquad (6.47)$$

For example, Eq. (6.47) holds for a harmonic process with a random initial phase (see Example 6.5).

Measuring the characteristics of random processes. If a random process is ergodic, its realization of a sufficient length is a "typical" representative of the entire ensemble. A sufficiently complete characterization of the entire process can be developed by analysing this only realization.

A device to measure the univariate probability density of a random process can be built as follows. The univariate probability density of an ergodic random process should be treated as a quantity proportional to the relative time during which the realization exists at a level between x and $x + \Delta x$. Suppose that we have a network with two inputs one of which accepts the realization under analysis, $x(t)$, and the other accepts a constant reference voltage whose level x_0 can be adjusted within certain limits. The output of the device delivers rectangular video pulses of constant amplitude, the start and finish of which are determined by the instants when the instantaneous value of the random signal corresponds to either x_0 or $x_0 + \Delta x$. If, now, we measure with an ordinary pointer-type instrument the average current produced by a train of video pulses, the meter indications will be proportional to the probability density $p(x_0)$ to within a constant.

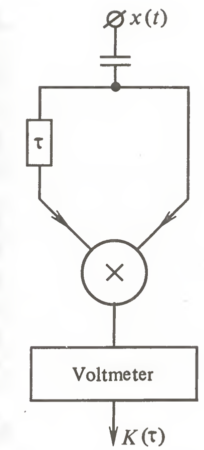


Measuring the probability density of a random process

Any sufficiently sluggish pointer-type instrument can be used to find the expectation (mean) of a random process [see Eq. (6.43)].

As follows from Eq. (6.44), the instrument used to measure the variance of a random process should have at its input a capacitor to block the passage of the d.c. component. The subsequent steps in the procedure, namely squaring and time-averaging, are usually performed by an ordinary, sufficiently sluggish square-law voltmeter.

The operating principle of the instrument measuring the autocorrelation function (called the *correlator* or the *correlation detector*) arises from Eq. (6.45). Here, after the d.c. component is filtered out, the instantaneous level of the random signal is divided between two channels one of which delays its share of the signal for a time τ. From the two channels, the divided signals are fed to a multiplier. The product thus formed is in turn applied to a sluggish element which produces an averaged output.



The correlator (correlation detector)

The above techniques of measuring the characteristics of random processes are based on analog operations. Of late, there has been a growing trend towards using digital instruments for the purpose. For their operation they depend on the sampling of the random signal in question and the processing of the resultant sampled numbers in accord with Eqs. (6.43) through (6.45).

Work Problem 15

The cross-correlation function of two random processes. In many cases it is of interest to know the extent to which two stationary random processes, X and Y, are statistically related to each other. For this purpose, we introduce the *cross-correlation* of the two processes

$$K_{xy}(t_1, t_2) = \overline{[x(t_1) - m_x][y(t_2) - m_y]}$$
$$K_{yx}(t_1, t_2) = \overline{[y(t_1) - m_y][x(t_2) - m_x]} \qquad (6.48)$$

Stationary related random processes

Random processes are called *stationary related* if the cross-correlations $K_{xy}(t_1, t_2)$ and $K_{yx}(t_1, t_2)$ depend only on the time difference $\tau = t_2 - t_1$, and not on t_1 and t_2 themselves. Then the cross-correlation becomes the *cross-correlation function** such that

$$K_{xy}(\tau) = K_{yx}(-\tau) \qquad (6.49)$$

Suppose that the random processes $X(t)$ and $Y(t)$ are statistically independent in the sense that for the instantaneous values $x = x(t)$ and $y_\tau = y(t+\tau)$ the bivariate probability density is

$$p(x, y_\tau) = p(x)\,p(y_\tau)$$

* Many authors call it the cross-covariance function.– *Translator's note.*

irrespective of the value of τ. Then

$$K_{xy}(\tau)=\int_{-\infty}^{\infty}(x-m_x)\,p(x)\,dx\int_{-\infty}^{\infty}(y_\tau-m_y)\,p(y_\tau)\,dy_\tau=0$$

or, in words, from the statistical independence it follows that the two processes are uncorrelated. In general, the converse is not true.

Stationary Gaussian random processes. These mathematical models of random signals are widely used in communication theory in order to describe the statistical phenomena due to a large number of independent random variables, that is, when the Central Limit Theorem applies. By definition, the nth-order multivariate probability density of a stationary Gaussian process depends on n–1 time arguments $\tau_i=t_i-t_1$, $i=2, 3, \ldots, n$ in the following way:

$$p(x_1,\ldots,x_n,\tau_1,\ldots,\tau_{n-1})=\frac{1}{\sigma^n(2\pi)^{n/2}D^{1/2}}$$

$$\times\exp\left[-\frac{1}{2D\sigma^2}\sum_{i,j=1}^{n}D_{ij}(x_i-m)(x_j-m)\right] \tag{6.50}$$

Here, the notation is the same as is used in Eq. (6.26). The elements of the correlation matrix of this random process can be found from the normalized autocorrelation matrix $R_{ij}=R(\tau_i-\tau_j)$.

Frequently, use is made of the bivariate Gaussian probability density

$$p(x_1,x_2,\tau)=\frac{1}{2\pi\sigma^2\sqrt{1-R^2(\tau)}}$$

$$\times\exp\left\{-\frac{(x_1-m)^2-2R(\tau)(x_1-m)(x_2-m)+(x_2-m)^2}{2\sigma^2[1-R^2(\tau)]}\right\} \tag{6.51}$$

Stationary Gaussian processes occupy an exceptional place among all the other random processes–any of the multivariate probability densities can be specified by only two characteristics: the expected value (mean) and the autocorrelation function. This is the reason why most of the results obtained in statistical communication theory have special reference to stationary Gaussian processes.

A. A. Markov (1856-1922), a prominent Russian mathematician

Markov processes. In concluding this overview of the general methods used to describe random signals, it is worth while to dwell on a rather special mathematical model widely used in communication theory. What we mean are the so-called *Markov random processes.*

Suppose that the state of a physical system can be characterized by one of the numbers belonging to a finite set $x_1, x_2, \ldots, x_N$. The change of state occurs in a random manner at fixed instants of time $t_1<t_2<t_3<\ldots$. Such a random process is called the *simple Markovian chain* if the probability of observing the system in the state x_i during the kth step solely depends on the state x_j in which the system resided during the previous, $(k-1)$st, step. The states associated with the steps $(k-2)$, $(k-3)$, ... are immaterial.

On designating this transition probability as p_{ij} $(i, j=1, \ldots, N)$, we may introduce the following matrix

$$\Pi=\begin{bmatrix} p_{11}\,p_{12}\cdots p_{1N}\\ \vdots \\ p_{N1}\,p_{N2}\cdots p_{NN}\end{bmatrix}$$

which completely describes the statistical properties of the Markovian chain.

The basic problem that has to be solved in the theory of Markovian chains may be stated as follows: Given the matrix Π and the initial state of the process, we are to find the probability that the system will reach some fixed state x_m in exactly n steps.

The idea of Markov random processes can be generalized to the continuous-time case such that the realizations can likewise take on a continuous set of values.

It is to be stressed that the Markovian properties of a random process are related to the dynamic behaviour of the associated physical system and are in no way indicative of the form of the probability density. Among other things, a Markov random process may or may not be a Gaussian process.

There is a large body of literature on the theory and applications of Markov random processes (see, for example, [3] and [14]).

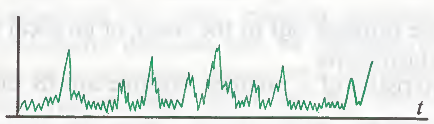
Summary

- Statistical relations are brought out in the study of physical systems formed by a large number of smaller subsystems.
- The principal characteristics of a random variable are its distribution function and its probability density.
- The numerical parameters used to describe a random variable are its moments, such as the expected value (or mean) and the variance.
- The statistical relations existing between the individual components of an nth-order (vector) random variable are defined in terms of joint second-order moments called correlation coefficients.
- Uncorrelated Gaussian variables are statistically independent.
- Under the Central Limit Theorem, in the limit $m\to\infty$ the sum of m independent random variables is normally (Gaussian) distributed.

- A random process is specified by an infinite ensemble of its realizations.
- The most important moment functions of a random process are its expected value (mean), variance, and autocorrelation function.
- A random process is called stationary if its statistical characteristics are time-invariant.
- The characteristics of stationary random processes which possess the property of ergodicity may be investigated experimentally by analysing only one realization.
- The expected value (mean) and the autocorrelation function make it possible to calculate any nth-order multivariate probability density of a stationary Gaussian random process.
- Realizations of Markov processes are chains of randomly changing states; the probability of observing any one state in a given step solely depends on the nearest previous state.

Review Questions

1. State the axioms of probability theory.
2. What is the difference between the population probability and the sample probability?
3. Name the basic properties of the probability density of a random variable.
4. Formulate the principle by which the average values of the quantities functionally related to random variables are found.
5. What is the procedure for finding the probability density of a function of a random variable in the case of a unique and a non-unique relationship?
6. How is the probability density of a random variable related to its characteristic function?
7. What is the meaning of correlation between two random variables?
8. Which requirement is more rigorous: the uncorrelatedness or the statistical independence of random variables?
9. List the salient features of an nth-order Gaussian random variable.
10. State the Central Limit Theorem.
11. What is the difference between a random process and a random realization?
12. An experiment has yielded the following realization of a random signal:



Can it belong to an ensemble of realizations of a Gaussian random process in principle? How valid is such a statement?

13. Define (a) a wide-sense stationary random process and (b) a narrow-sense stationary random process.
14. What is the principal property of an ergodic random process?
15. What is the physical meaning of the variance of an ergodic random process?
16. Define the cross-correlation function of two random processes.

Problems

1. When a message is transmitted over a communication channel, an average of 0.5% symbols are received in error. The message transmitted is 120 symbols long. What is the probability of faithfully reproducing the transmitted message?

2. The probability density of a random variable X is

$$p(x) = a\exp(-b|x|)$$

Find the relation between the numbers a and b, stemming from the condition of normality.

3. A random variable X is uniformly distributed within the interval (0, 5). The probability of detecting this variable at the ends of the interval is the same and equal to 0.3. Calculate the distribution function and the probability density for the random variable in question and construct the applicable plots.

4. Find the mean and variance for the random variable of Problem 3.

5. Find the mean and variance for a random variable whose probability density is

$$p(x) = (\alpha/2)\exp(-\alpha|x|)$$

for $\alpha > 0$.

6. The characteristic function $\Theta(v)$ of a random variable X has the form

$$\Theta(v) = 1/(1+v^2)$$

Find the probability density $p(x)$ of this random variable.

7. Find the relation between the probability density $p(x)$ of a random variable X and the probability density $p(y)$ of another random variable Y obtained by the following functional transformation:

$$y = \exp(-x^2)$$

8. The joint probability density $p(x_1, x_2)$ of a second-order random variable has the form

$$p(x_1, x_2) = (a^2/\pi)\exp[-a^2(x_1^2 + x_2^2)]$$

Find the probability densities, means and variances of x_1 and x_2.

9. Prove that for a wide-sense stationarity of a random process $X(t)$ with realizations

$$x(t) = A\cos\omega t + B\sin\omega t$$

it is necessary and sufficient for the random variables A and B to have the following properties:

(a) $\bar{A} = \bar{B} = 0$;

(b) $\sigma_A^2 = \sigma_B^2$;

(c) $\overline{AB} = 0$

10. A random process $Z(t)$ is the sum of two Gaussian random processes $X(t)$ and $Y(t)$ which have time-invariant means m_x and m_y and time-invariant variances σ_x^2 and σ_y^2, respectively. Find the univariate probability density of the total process.

Advanced Problems

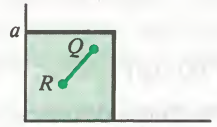
11. There is a signal which is the sum of harmonic waves of the same frequency. The components have the same amplitude equal to 5 V. The initial phases of the components may take on independently only two values, 0° and 180°. The number of components is 30. Find the probability that the resultant amplitude will exceed 40 V.

12. Prove that if a random variable Z is the sum of two independent random variables X and Y, then its probability density is the convolution of the individual (marginal) densities:

$$p_Z(z) = \int^{\infty} p_Y(\zeta)\, p_X(z-\zeta)\, d\zeta$$

13. A random point $Q(x_1, x_2)$ is uniformly

distributed in a square of side a:



Find the mean and variance for the length of the random segment RQ joining the point to the centre of the square.

14. The coordinates (x, y) of a random point in a plane are independent Gaussian random variables of mean value $m_x = m_y = 0$ and of variance $\sigma_x^2 = \sigma_y^2 = \sigma^2$. Find the probability density of the length of the random position vector of the point.

15. Consider the feasibility of an instrument for measuring the bivariate probability density of an ergodic random process.

Chapter 7

The Correlation Theory of Random Processes

In addition to a complete description of the properties of random signals with the aid of multivariate probability densities, a simpler approach may be taken in which random processes are characterized by their moment functions. The theory of random processes based on the use of at most second moment functions has come to be known as *correlation theory*. In this chapter we set out to demonstrate that there is a deep and close relation between the correlation and spectral properties of random signals.

Correlation theory

7.1 The Spectral Representation of Stationary Random Processes

In Chap. 2, the spectral theory has been set forth for deterministic signals. Since individual realizations are probabilistic in their behaviour, the methods of spectral analysis cannot be carried over to the theory of random processes directly. However, a number of important spectral characteristics of random waves can be derived by taking the Fourier transforms of some functions formed by averaging the realizations.

The spectrum of realizations. Consider a stationary random process $X(t)$ of mean value zero, $\overline{x} = 0$. A single realization of this process is a deterministic function which can be represented by the following spectral expansion:

$$x(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} S(\omega) \exp(j\omega t)\, d\omega \qquad (7.1)$$

with a certain deterministic spectrum $S(\omega)$.

In order to describe the entire ensemble of realizations that form the process $X(t)$, we must assume that the $S(\omega)$'s are themselves random functions of frequency. Thus, a random process in the time domain is connected to another random process in the frequency domain. There is a one-to-one correspondence between the individual realizations of the two processes. If the realizations of a random process are represented as in Eq. (7.1), the process is said to appear in a *spectral representation.*

Spectral representation of a random process

The crucial question in the theory of random processes is: *What properties should the random functions $S(\omega)$ possess for the process $X(t)$ to be stationary?*

Properties of a random spectrum. In order to answer the above question, we will define the mean of the instantaneous values over

the ensemble:

$$\overline{x}=\frac{1}{2\pi}\int_{-\infty}^{\infty}\overline{S(\omega)}\exp(j\omega t)\,d\omega=0$$

The above equality will be satisfied identically for any t, if we require that

$$\overline{S(\omega)}=0 \tag{7.2}$$

In other words, the random spectrum of individual realizations of a stationary random process must have a mean value of zero at all frequencies.

Now we should determine the conditions under which the autocorrelation function $K(\tau)$ will depend solely on the time shift τ between the sections. Since $x(t)$ is a real signal, we may use, in addition to Eq. (7.1), also the equality

For a complex process, its real and imaginary parts must be considered separately

$$x(t)=x^*(t)=\frac{1}{2\pi}\int_{-\infty}^{\infty}S^*(\omega)\exp(-j\omega t)\,d\omega \tag{7.3}$$

Using the spectral expansion of the random signal, we may write its autocorrelation function as

$$\begin{aligned}K(\tau)&=\overline{x(t)x(t+\tau)}=\overline{x^*(t)x(t+\tau)}\\&=\frac{1}{(2\pi)^2}\iint_{-\infty}^{\infty}\overline{S(\omega)S^*(\omega')}\exp(j\omega\tau)\\&\quad\times\exp[j(\omega-\omega')t]\,d\omega\,d\omega'\\&=\frac{1}{(2\pi)^2}\int_{-\infty}^{\infty}\exp(j\omega\tau)\,d\omega\int_{-\infty}^{\infty}\overline{S(\omega)S^*(\omega')}\\&\quad\times\exp[j(\omega-\omega')t]\,d\omega'\end{aligned} \tag{7.4}$$

Note that $S(\omega)$ takes on complex values

Here the inner integrand contains the factor $\overline{S(\omega)S^*(\omega')}$ which has the meaning of the autocorrelation function of a random spectrum. For $K(\tau)$ not to be dependent on t, it is necessary, as is seen from Eq. (7.4), to require that the following proportionality be satisfied:

$$\overline{S(\omega)S^*(\omega')}\sim\delta(\omega-\omega') \tag{7.5}$$

Delta correlation

Thus, the random spectrum $S(\omega)$ of a stationary random process has a very specific structure: The spectra corresponding to any two non-coincident frequencies are mutually uncorrelated, whereas the mean square (variance) of the random spectrum is large without bound. This form of statistical association is called *delta correlation.*

The power spectrum of a stationary random process. Let us introduce a frequency-dependent proportionality factor in Eq. (7.5) and re-write the equation as follows:

$$\overline{S(\omega)S^*(\omega')}=2\pi W(\omega)\delta(\omega-\omega') \tag{7.6}$$

The function $W(\omega)$, playing a fundamental role in the theory of stationary random processes, is called the *power spectrum** of the process $X(t)$.

The power spectrum of a random process

On substituting (7.6) into (7.4), we obtain an important result:

$$K(\tau)=\frac{1}{2\pi}\int_{-\infty}^{\infty}W(\omega)\exp(j\omega\tau)\,d\omega \tag{7.7}$$

Thus *the autocorrelation function and the power spectrum of a stationary random process of mean value zero are each other's Fourier transforms.* This relation is also called the *Wiener-Khinchin (W-K) theorem.*

Hence,

$$W(\omega)=\int_{-\infty}^{\infty}K(\tau)\exp(-j\omega\tau)\,d\tau \tag{7.8}$$

To clarify the physical significance of the power spectrum, let us set $\tau=0$ in Eq. (7.7). Then, since $K(0)=\sigma^2$, we get

$$\sigma^2=\frac{1}{2\pi}\int_{-\infty}^{\infty}W(\omega)\,d\omega \tag{7.9}$$

The variance σ^2, equal to the mean power in the stationary random process, is thus the sum of the contributions from all frequencies. The quantity $W(\omega)$ is a measure of the mean power in the process per unit of bandwidth in the vicinity of the chosen frequency ω.

If a random process is voltage, its power spectrum has the dimensions of $V^2\,s\,rad^{-1}$

Physically, the power spectrum is real and non-negative:

$$W(\omega)\geqslant 0 \tag{7.10}$$

This property imposes rather rigorous constraints on the form of the permissible autocorrelation functions. (We have already run

* The alternative terms are the *power density spectrum* (*Handbook of Automation, Control and Computation*, Ed. Eu. M. Grabbe, S. Ramo, and D. E. Wooldridge. New York: John Wiley and Sons, Inc., 1958), the *spectral density* (C. W. Helstrom, *Statistical Theory of Signal Detection*. Oxford: Pergamon Press, 1968), the *power spectral density* (L. E. Franks, *Signal Theory*. Englewood Cliffs, New Jersey: Prentice Hall Inc., 1969).– *Translator's note.*

into a similar situation in Chap. 3 in connection with the correlation properties of deterministic signals.)

Solve Problem 1

One more important point should be stressed. The power spectrum of a stationary random process, being always real, does not carry any information about the phase relations between the individual spectral components. Therefore, it is impossible in principle to reconstruct any individual realization of a random process from its power spectrum.

The one-sided power spectrum. Since $K(\tau)$ is an even function of τ, the corresponding spectrum $W(\omega)$ is likewise an even function of the frequency ω. Therefore, the Fourier transform pair, (7.7) and (7.8), can be written, using only the integrals between semi-infinite limits:

$$K(\tau) = \frac{1}{\pi}\int_0^\infty W(\omega)\cos\omega\tau\, d\omega \tag{7.11}$$

$$W(\omega) = 2\int_0^\infty K(\tau)\cos\omega\tau\, d\tau \tag{7.12}$$

Solve Problem 2

It is expedient to introduce the so-called *one-sided power spectrum* $F(\omega)$ of a random process, by defining it as follows:

$$F(\omega) = \begin{cases} W(\omega)/\pi & \text{for } \omega > 0 \\ 0 & \text{for } \omega < 0 \end{cases} \tag{7.13}$$

The function $F(\omega)$ makes it possible to express the variance of a stationary random process as an integral over positive (entirely real or physical) frequencies:

$$\sigma^2 = K(0) = \int_0^\infty F(\omega)\, d\omega \tag{7.14}$$

The one-sided power spectrum may have the dimensions of, say, $V^2\ Hz^{-1}$

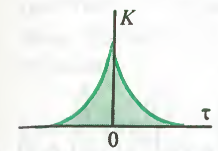
By the same token, we may define the one-sided power spectrum $F(f)$, which is the mean power in the process per unit frequency interval (1 Hz):

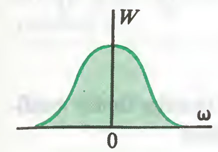
$$F(f) = \begin{cases} 2W(2\pi f) & \text{for } f > 0 \\ 0 & \text{for } f < 0 \end{cases}$$

such that

$$\sigma^2 = \int_0^\infty F(f)\, df$$

The W-K theorem is a most important tool in the applied theory of random processes and comes in useful when handling a large variety of problems. Consider some of the most typical examples.





Example 7.1. *A random process with an exponential autocorrelation function.*

Suppose we know that the process $X(t)$ has an autocorrelation function of the form

$$K(\tau) = \sigma^2\exp(-\alpha|\tau|)$$

in which α is some real and positive parameter. On the basis of Eq. (7.12), the associated power spectrum is

$$W(\omega) = 2\sigma^2\int_0^\infty \exp(-\alpha\tau)\cos\omega\tau\, d\tau = 2\alpha\sigma^2/(\alpha^2+\omega^2)$$

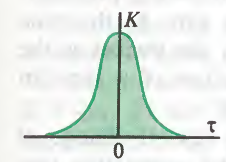
Hence, the one-sided power spectrum is

$$F(\omega) = (2/\pi)\,\alpha\sigma^2/(\alpha^2+\omega^2)$$

From the accompanying plot it is seen that the power spectrum of the process in question has a well-pronounced low-frequency character—the power spectrum is a maximum at frequency zero.

Example 7.2. *Suppose that the power spectrum of a random process $X(t)$ is described by a Gaussian function (a quadratic exponential function)*

$$W(\omega) = W_0\exp(-\beta\omega^2)$$



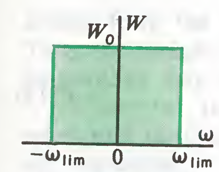
Let us find the autocorrelation function by using Eq. (7.11):

$$K(\tau) = (W_0/\pi)\int_0^\infty \exp(-\beta\omega^2)\cos\omega\tau\, d\omega = (W_0/2)\sqrt{(1/\pi\beta)}\exp(-\tau^2/4\beta)$$

Thus, a Gaussian power spectrum leads to a Gaussian autocorrelation function as well.

The variance of the random process in question is

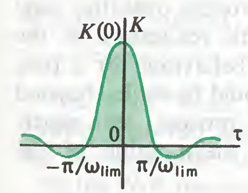
$$\sigma^2 = W_0/2\sqrt{\pi\beta}$$



Example 7.3. *A stationary random process with a band-limited low-frequency power spectrum.*

Let the process $X(t)$ be characterized by a power spectrum of the form

$$W(\omega) = \begin{cases} W_0 & \text{for } -\omega_{lim} < \omega < \omega_{lim} \\ 0 & \text{elsewhere} \end{cases}$$



Using Eq. (7.11), we find the autocorrelation function to be

$$K(\tau) = (W_0/\pi)\int_0^{\omega_{lim}} \cos\omega\tau\, d\omega = (W_0\,\omega_{lim}/\pi)(\sin\omega_{lim}\tau/\omega_{lim}\tau)$$

The variance of the process is

$$\sigma^2 = K(0) = W_0\,\omega_{lim}/\pi$$

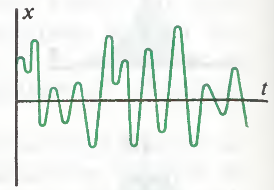
In this case, it is convenient to use the one-sided power spectrum

$$F(\omega)=\begin{cases} F_0=(W_0/\pi) \text{ for } 0<\omega<\omega_{\lim} \\ 0 \text{ elsewhere} \end{cases}$$

which makes it possible to write the expression for the variance as an easy-to-memorize product of the power spectrum by the frequency band occupied by the signal:

$$\sigma^2 = F_0\,\omega_{\lim}$$

It is important to note that the autocorrelation function of the process in question takes alternate signs, the change of sign occurring at each shift τ which is a multiple of $\pi/\omega_{\lim}$. In other words, as τ increases, the mean of the product $x(t)x(t+\tau)$ will first be positive, then negative, then again positive, and so on. This behaviour of the autocorrelation function points to the *quasi-periodicity* of any realization of the random process (of course in a probabilistic rather than absolute sense).

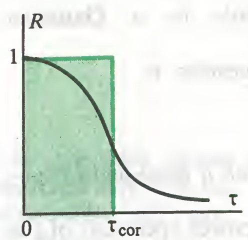


A quasiperiodic realization

From a physical point of view, the quasi-periodic realizations of random processes occur when the random waves are sums of a large number of independent r.f. pulses.

The correlation time. In general, the random processes of concern to statistical communication theory have the following property: Their autocorrelation function $K(\tau)$ approaches zero as the time shift τ increases. The faster the decrease in $K(\tau)$, the weaker is the statistical relation between the instantaneous values of a random signal observed at two different instants of time.

It is convenient to evaluate numerically the rate of change of realizations of a random process in terms of the *correlation time* τ_{cor}, defined as



The two shapes are the same in area

$$\tau_{cor}=\frac{1}{K(0)}\int_0^\infty K(\tau)\,d\tau=\int_0^\infty R(\tau)\,d\tau \qquad (7.15)$$

Roughly speaking, we can predict in probabilistic terms the behaviour of any realization of a random process over the time interval τ_{cor}, if we know the behaviour of this realization in the "past". Any attempt, however, to predict this behaviour for a time substantially exceeding the correlation time would be futile–beyond that interval isntantaneous values of the process are nearly independent random variables, that is, the mean value of the product $x(t)x(t+\tau)$ is close to zero.

▲ **Solve Problem 5**

The effective bandwidth. Let the random process in question be characterized by a one-sided power spectrum, $F(\omega)$, such that F_{max} is the extremal value of the function. Mentally we may replace

this random process with any other process in which the power spectrum is constant and equal to F_{max} within the frequency band $\Delta\omega_{eff}$ chosen such that the mean power is the same in both processes:

$$F_{max}\Delta\omega_{eff}=\int_0^\infty F(\omega)\,d\omega$$

Hence the expression for the effective bandwidth of a random process is

$$\Delta\omega_{eff}=(1/F_{max})\int_0^\infty F(\omega)\,d\omega \qquad (7.16)$$

This numerical characteristic is frequently used in engineering calculations as it provides a ready means for finding the variance of a noise voltage:

$$\sigma^2=F_{max}\Delta\omega_{eff}$$

For example, if we know that $F_{max}=5\times10^{-9}\,V^2$ s and $\Delta\omega_{eff}=3\times10^5\,s^{-1}$, then $\sigma^2=1.5\times10^{-3}\,V^2$. Hence, the rms noise voltage is $\sigma=\sqrt{\sigma^2}=39$ mV.

Obviously, the effective bandwidth of a random process can be specified in more than one way. For example, it may be defined as the interval at whose boundaries the power spectrum decreases to $0.1F_{max}$. In any case, the correlation time τ_{cor} and the effective bandwidth $\Delta\omega_{eff}$ will be connected by the uncertainty relation

▲ **Solve Problem 4**

$$\Delta\omega_{eff}\tau_{cor}=O(1)$$

arising from the properties of Fourier transforms (see Chap. 2). Thus, the broader the bandwidth of noise, the more chaotically its realizations vary with time.

White noise. In communication theory, the term "white noise" refers to a stationary process whose power spectrum is the same at all frequencies:

$$W(\omega)=W_0=\text{const}$$

The term "white noise" suggestively stresses the analogy with "white" (natural) light in which all the spectral components within the visible region have about the same intensity.

● **White noise**

By the W-K theorem, the autocorrelation function of white noise

$$K(\tau)=(W_0/2\pi)\int_{-\infty}^{\infty}\exp(j\omega\tau)\,d\omega=W_0\delta(\tau) \qquad (7.17)$$

which is an indication of an infinitely large mean power in white noise.

Alternatively, white noise is referred to as a delta-correlated random process. The uncorrelatedness of instantaneous values of its realizations implies that they vary with time at an infinitely high rate: However small the time interval τ may be, the instantaneous value of the signal can change over that time by any value we may like to specify in advance.

White noise is an abstract mathematical model, and a physical process answering this model is, of course, non-existent in nature. This does not prevent, however, using it as an approximation for sufficiently broadband real random processes in cases where the passband of the circuit responding to the signal is substantially narrower than the effective bandwidth of the noise.

7.2 Differentiation and Integration of Random Processes

In this section, we will be concerned with the properties displayed by realizations of random processes subjected to differentiation and integration. It will be shown that the most important characteristic defining the differentiability of a random process is its autocorrelation function.

Probabilistic treatment of convergence and continuity. In the theory of random processes we have somewhat to extend the classical concept of the convergence of a sequence of numbers towards their limit. Thus, if $\{x_n\}$ is a random sequence, then it is not always that in the limit $m, n \to \infty$ the difference $|x_m - x_n|$ should be less than any predetermined number, however small.

A random sequence $\{x_n\}$ is said to converge towards a nonrandom number x *in probability* if, for any $\varepsilon > 0$,

Convergence in probability

$$\lim_{n\to\infty} P(|x_n - x| > \varepsilon) = 0 \tag{7.18}$$

The requirement for convergence in probability, applied in all problems involving random processes, is less rigorous than the classical criteria for the convergence of deterministic sequences.

The property of continuity for a random process is defined in a similar way. A random process $X(t)$ is said to be *continuous* at point $t = t_0$, if the following limiting equality is satisfied

Continuity of a random process

$$\lim_{t_1\to t_0} \overline{[x(t_1) - x(t_0)]^2} = 0 \tag{7.19}$$

The derivative of a random process. Suppose that any realization $x(t)$ of a random process $X(t)$ can be applied to a differentiating network which produces at its output a new realization $y(t) = \mathrm{d}x/\mathrm{d}t$. The collection of realizations $y(t)$ forms a random process $Y(t)$ called the *derivative* of the process $X(t)$. Symbolically, this is stated as follows:

$$Y(t) = \mathrm{d}X/\mathrm{d}t$$

Let $X(t)$ be a stationary random process of a known mean value $\bar{x} = m_x$. In order to find the mean of the derivative, we should take the average over realizations:

$$m_y = \bar{y} = \overline{\mathrm{d}x/\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}t} m_x = 0 \tag{7.20}$$

To sum up, the differentiation of any stationary random signal produces a new random signal of mean value zero.

Finding the autocorrelation of a derivative poses a more difficult problem. Without loss of generality, let us suppose that the mean value of the original process is $m_x = 0$. (If this should be not the case, we may always choose a new process $\tilde{X}(t)$ whose realizations are $\tilde{x}(t) = x(t) - m_x$.) Taking advantage of the fact that

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \lim_{\Delta t\to 0} \frac{x(t+\Delta t) - x(t)}{\Delta t}$$

we may write the autocorrelation function of the derivative as

$$K_y(\tau) = \overline{y(t)\,y(t+\tau)}$$

$$= \lim_{\Delta t\to 0} \overline{\frac{x(t+\Delta t) - x(t)}{\Delta t}\;\frac{x(t+\tau+\Delta t) - x(t+\tau)}{\Delta t}}$$

$$= \lim_{\Delta t\to 0} \frac{1}{(\Delta t)^2}\Big[\overline{x(t+\Delta t)\,x(t+\tau+\Delta t)}$$

$$- \overline{x(t+\Delta t)\,x(t+\tau)} - \overline{x(t)\,x(t+\tau+\Delta t)} + \overline{x(t)\,x(t+\tau)}\Big]$$

Work Problem 7

The four terms within the square brackets are the autocorrelation functions of the original process, as found for various values of time delay. It is an easy matter to see that

$$K_y(\tau) = \lim_{\Delta t\to 0} \frac{1}{(\Delta t)^2}\,[2K_x(\tau) - K_x(\tau - \Delta t) - K_x(\tau + \Delta t)]$$

The right-hand side of the above equality can be recognized as a finite-difference representation of the second derivative of $K_x(\tau)$, taken with the opposite sign. This leads us to an important formula

$$K_y(\tau) = -K_x''(\tau) = -\sigma_x^2 R''(\tau) \tag{7.21}$$

Differentiable and nondifferentiable random processes. By definition, a random process $X(t)$ is *differentiable* if its derivative has a finite variance.

In accord with Eq. (7.21), the variance of a derivative is

$$\sigma_y^2 = -K_x''(0) = -\sigma_x^2 R''(0)$$

The condition for the differentiability of a random process

Therefore, for a random process to be differentiable it is necessary that the second derivative of its autocorrelation function be finite at zero-crossing and, in consequence, the first derivative be continuous at that point.

Nondifferentiable random processes are those for which the autocorrelation function is of the form $\sigma^2 \exp(-\alpha|\tau|)$, considered in Example 7.1. On differentiating this function, it can be easily seen that at zero-crossing the derivative changes in magnitude stepwise by $-2\sigma^2\alpha$.

In communication theory one has often to deal with random processes whose autocorrelation functions have the form

$$K(\tau) = \sigma^2(1+\alpha|\tau|)\exp(-\alpha|\tau|) \tag{7.22}$$

The first derivative of the above function

$$K'(\tau) = \begin{cases} -\alpha^2\sigma^2\tau\exp(-\alpha\tau) & \text{for } \tau > 0 \\ -\alpha^2\sigma^2\tau\exp(\alpha\tau) & \text{for } \tau < 0 \end{cases}$$

is continuous at zero-crossing, so the autocorrelation function in (7.22) is associated with a differentiable process. Any real random signals are always sufficiently continuous, that is, they are differentiable. In theoretical studies, however, use is often made of mathematical models corresponding to nondifferentiable processes. As a rule, this is the case when realizations of a random process are formed from a large number of independent variables. Although the contribution by one such variable (say, the current pulse produced by the motion of a single electron) is negligible, precisely these components determine the "fine structure" of the realization. In consequence, realizations of such a process can, in the limit, take the form of a function everywhere continuous, but nowhere differentiable.

Solve Problem 8

The power spectrum of a derivative. We set out to establish the relation between the power spectra of an original process and of its derivative. Let $X(t) \leftrightarrow W_x(\omega)$. By the W-K theorem, the autocorrelation function of the original process is

$$K_x(\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty} W_x(\omega)\exp(j\omega\tau)\,d\omega$$

On the basis of (7.21), the autocorrelation function of the

derivative is

$$K_y(\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \omega^2 W_x(\omega)\exp(j\omega\tau)\,d\omega$$

Hence, the sought-for relation is

$$W_y(\omega) = \omega^2 W_x(\omega) \tag{7.23}$$

Resort is made to the differentiation of an integral with respect to the parameter

It is worth noting that the low-frequency components in the power spectrum of the derivative are weakened and the high-frequency components are boosted. On the basis of Eq. (7.23) we can say whether the process $X(t)$ is differentiable from the properties of its power spectrum: The variance of the derivative will be finite, if the integral

$$\sigma_y^2 = \frac{1}{2\pi}\int_{-\infty}^{\infty} \omega^2 W_x(\omega)\,d\omega < +\infty$$

exists.

For example, in the case of a random process having a low-frequency power spectrum (see Example 7.3), the variance of the derivative is

$$\sigma_y^2 = (W_0/2\pi)\int_{-\omega_{\lim}}^{\omega_{\lim}} \omega^2\,d\omega = W_0\omega_{\lim}^3/3\pi$$

Therefore, the process involved is differentiable.

Correlation between a random process and its derivative. In many communication-theory problems it is interesting to know the extent to which a random signal and its derivative are related to each other statistically. In order to answer this question, let us find the cross-correlation function $K_{xy}(\tau)$ of the random processes $X(t)$ and $Y(t) = dX/dt$, assuming that both are stationary and of mean value zero. Then

$$K_{xy}(\tau) = \overline{x(t)\,y(t+\tau)} = \overline{x(t)\frac{d}{d\tau}x(t+\tau)} = \frac{d}{d\tau}\overline{x(t)\,x(t+\tau)}$$

Hence,

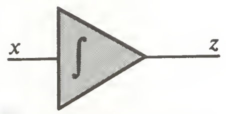
$$K_{xy}(\tau) = dK_x(\tau)/d\tau \tag{7.24}$$

As will be recalled, the function $K_x(\tau)$ shows even symmetry. Therefore, for $\tau = 0$, the derivative $dK_x/d\tau$ vanishes always. On the basis of Eq. (7.24), we may conclude that the *instantaneous values of a signal and of its derivative, observed at the same instant of time, are uncorrelated random variables.* A stronger assertion can be made with regard to Gaussian random processes. Here, a random signal and its derivative are *statistically independent.*

A random process and its derivative are uncorrelated

The integral of a random process. Let the random process $Z(t)$

be called the variable-upper-limit *definite integral* of a random process $X(t)$ if the following correspondence exists between their realizations $z(t)$ and $x(t)$:



The integrator

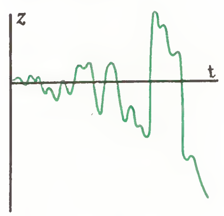
$$z(t)=\int_0^t x(t_1)\,dt_1 \tag{7.25}$$

The physical significance of this is that the signals $z(t)$ are observed at the output of an ideal integrator, with the input signals $x(t)$ commencing to arrive at time zero.

If the process $X(t)$ is stationary and of mean value m_x, the mean value of the signal at the output of the integrator will be

The input random signal

and the output random signal



of an integrator

$$m_z=\bar{z}(t)=\int_0^t \bar{x}(t_1)\,dt_1=m_x t \tag{7.26}$$

Thus, the condition $m_x\neq 0$ immediately leads to the nonstationarity of the random process $Z(t)$.

However, even if the input signal is of mean value zero, the signal at the output of an ideal integrator will be a realization of a nonstationary random process. To prove this, let us find the autocorrelation function of the integral:

$$K_z(t_1,\ t_2)=\overline{\int_0^{t_1}\int_0^{t_2} x(t')x(t'')\,dt'dt''}$$

$$=\int_0^{t_1}\int_0^{t_2}\overline{x(t')x(t'')}\,dt'dt''=\int_0^{t_1}\int_0^{t_2}K_x(t',\ t'')\,dt'dt''$$

If the process $X(t)$ is stationary, the argument of the autocorrelation function under the integral in the above equation will be the time difference $t''-t'$, so

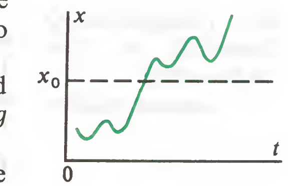
$$K_z(t_1,\ t_2)=\int_0^{t_1}\int_0^{t_2}K_x(t''-t')\,dt'\,dt'' \tag{7.27}$$

Since the right-hand side of Eq. (7.27) depends directly on t_1 and t_2, rather than on their difference, the signal at the integrator output cannot be a stationary random process.

The nonstationarity of the integral of a random process has an important physical significance – it implies that the level of fluctuations at the output of an ideal integrator rises without bound due to their accumulation.

Similar situations are frequently encountered in various fields of physics. An example is the well-known one-dimensional random walk of a particle (the Brownian motion) [14]. Here, the imaginary particle, on being equiprobably hit in two opposite directions, remains on the average in the same place, but its deviation from the mean position increases without bound in time.

The crossing problem. In statistical communication theory, the following problem related to the differential properties of random processes is of special interest. Suppose that realizations of some random process $X(t)$ are sufficiently continuous functions of time. It is required to determine how frequently these realizations cross some fixed level x_0. This problem naturally arises in, say, the noise-immunity analysis of communication systems subjected to fluctuation or random pulse noise.



An upward crossing

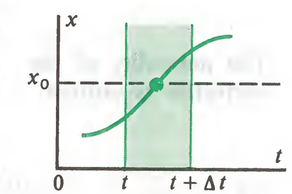
An event consisting in that a realization $x(t)$ crosses the specified level x_0 in the upward direction can be called the *upward crossing* of the process $X(t)$ at the level x_0.

Let us solve a very simple problem – we seek to find the average number of upward crossings per unit time. To this end, we mentally select on the t-axis a short interval of length or duration Δt. Assuming that the process $X(t)$ is stationary and continuous, it is always possible to specify so small an interval Δt that no or only one upward crossing will occur within it.

To begin with, we find the probability of the elementary event consisting in that in time Δt one upward crossing takes place. Obviously, it can occur if:

(a) $x(t)<x_0$ and (b) $x(t+\Delta t)>x_0$

Since



$$x(t+\Delta t)\approx x(t)+x'\Delta t$$

the condition (b) implies that

$$x_0-x'\Delta t<x(t)$$

Thus, a single upward crossing inside the interval Δt will take place, if the realization of the random process in that interval has a positive derivative, or an upward slope ($x'\geqslant 0$) and satisfies the inequality

$$x_0-x'\Delta t<x(t)<x_0$$

The probability P of the above event is easy to find if we know the joint bivariate probability density $p(x,\ x')$ of the realization and of its derivative, related to the same instant of time:

$$P=\int_0^\infty dx'\int_{x_0-x'\Delta t}^{x_0} p(x,\ x')\ dx=\Delta t\int_0^\infty p(x_0,\ x')x'\,dx' \tag{7.28}$$

The direct proportionality between the probability in (7.28) and the duration of the time interval Δt is an indication that the quantity $n(x_0)$, the average number of upward crossings per second

(or the crossing rate), can be defined by

$$n(x_0) = \int_0^\infty p(x_0, x') x' \, dx' \tag{7.29}$$

Crossings of Gaussian processes. The average number of upward crossings is far simpler to find if the process $X(t)$ is Gaussian. Then the instantaneous values of a realization and of its derivative observed at the same instants of time are statistically independent, that is,

$$p(x_0, x') = p(x_0) p(x') \tag{7.30}$$

Also, the derivative, obtained as a result of linear operations on the original process, is likewise normal. On combining (7.29) and (7.30), we get

$$n(x_0) = p(x_0) \int_0^\infty p(x') x' \, dx' \tag{7.31}$$

Suppose that we know the autocorrelation function

$K_x(\tau) = \sigma_x^2 R(\tau)$

of the process. Then the variance of the derivative will be

$\sigma_{x'}^2 = -\sigma_x^2 R''(0)$

The normality of the derivative is utilized

Hence, the probability density of the derivative is

$$p(x') = \frac{1}{\sqrt{2\pi}\sigma_x \sqrt{-R''(0)}} \exp\left\{-\frac{x'^2}{2\sigma_x^2[-R''(0)]}\right\} \tag{7.32}$$

By simple manipulations, we get

$$\int_0^\infty p(x') x' \, dx' = \frac{1}{\sqrt{2\pi}} \sigma_x \sqrt{-R''(0)}$$

On substituting the above expression into (7.31), the average number of upward crossings per second for a stationary Gaussian process is found to be

$$n(x_0) = (1/2\pi)\sqrt{-R''(0)} \exp(-x_0^2/2\sigma_x^2) \tag{7.33}$$

Quasi-frequency

The quasi-frequency of a stationary random process. In Sec. 7.1 it has been noted that some forms of random processes vary in time quasi-periodically. The numerical measure defining the rate of such variations is the *quasi-frequency* defined as the average number of zero-crossings per unit time. By virtue of Eq. (7.33), the quasi-frequency for a Gaussian process

$$n(0) = \frac{1}{2\pi}\sqrt{-R''(0)} \tag{7.34}$$

is wholly defined by the behaviour of the autocorrelation function at zero. Since

Quasi-frequency can only be determined for a differentiable random process

$-R''(0) = \sigma_{x'}^2/\sigma_x^2$

and the variance of the derivative is expressed in terms of the one-sided power spectrum of the process $X(t)$:

$$\sigma_{x'}^2 = \int_0^\infty \omega^2 F_x(\omega) \, d\omega$$

then the quasi-frequency may be re-cast in a form equivalent to that of Eq. (7.34):

$$n(0) = \frac{1}{2\pi\sigma_x}\left[\int_0^\infty \omega^2 F_x(\omega) \, d\omega\right]^{1/2} \tag{7.35}$$

Solve Problem 9

Example 7.4. *The quasi-frequency of a stationary Gaussian process with a band-limited low-frequency power spectrum* (see Example 7.3). Here,

$$\int_0^\infty \omega^2 F_x(\omega) \, d\omega = F_0 \omega_{\lim}^3/3$$

$$\sigma_x = \sqrt{F_0 \omega_{\lim}}$$

On substituting the above expressions into Eq. (7.35), we get

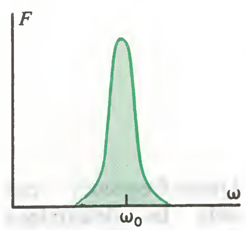
$n(0) = \omega_{\lim}/2\pi\sqrt{3} = f_{\lim}/\sqrt{3}$

The final result cannot be found directly.

7.3 Narrowband Random Processes

An extremely important role in telecommunications is played by a special class of random processes a salient feature of which is that their power spectrum has a sharply defined peak around a certain nonzero frequency. What follows is an analysis of the statistical properties of such narrowband random processes. The exposition is limited to the case of Gaussian processes most frequently encountered in practice. Also, it is for Gaussian processes that a number of important results can be obtained within the framework of correlation theory.

The autocorrelation function of a narrowband random process.



The power spectrum of a narrowband random process

We consider a random process $X(t)$ whose one-sided power spectrum $F(\omega)$ is concentrated in the vicinity of an arbitrarily chosen frequency ω_0.

By the W-K theorem, the autocorrelation function of the process in question is

$$K(\tau)=\int_0^{\infty} F(\omega)\cos\omega\tau\,d\omega \tag{7.36}$$

Let us mentally shift the power spectrum of the process from the vicinity of frequency ω_0 into the vicinity of zero frequency by a change of variable, $\omega=\omega_0+\Omega$. Then Eq. (7.36) takes on the following form:

$$K(\tau)=\int_{-\omega_0}^{\infty} F(\omega_0+\Omega)\cos[(\omega_0+\Omega)\tau]\,d\Omega \tag{7.37}$$

Since we have assumed that the process $X(t)$ is narrowband, its power spectrum $F(\omega)$ is vanishingly small at frequencies close to zero. Therefore, we may, to a high degree of accuracy, replace the lower integration limit in Eq. (7.37) with $-\infty$ and re-write the autocorrelation function as

$$K(\tau)=a(\tau)\cos\omega_0\tau-b(\tau)\sin\omega_0\tau \tag{7.38}$$

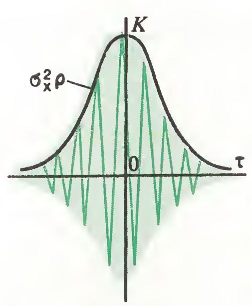
where

$$a(\tau)=\int_{-\infty}^{\infty} F(\omega_0+\Omega)\cos\Omega\tau\,d\Omega$$

and

$$b(\tau)=\int_{-\infty}^{\infty} F(\omega_0+\Omega)\sin\Omega\tau\,d\Omega$$

are slowly varying functions of the argument τ. It is to be noted that $a(\tau)$ is even and $b(\tau)$ is odd about τ.



The autocorrelation function of a narrowband random process

The autocorrelation function of a narrowband random process is especially simple and straightforward when the power spectrum $F(\omega)$ is symmetrical about its centre frequency ω_0. Then $b(\tau)=0$ and

$$K(\tau)=a(\tau)\cos\omega_0\tau \tag{7.39}$$

The factor $a(\tau)$ plays the role of an envelope which varies more slowly than the factor $\cos\omega_0\tau$. Frequently it is convenient to invoke a normalized envelope, $\rho(\tau)$, for the autocorrelation function of a narrowband random process, by defining it as

$$a(\tau)=\sigma_x^2\rho(\tau)$$

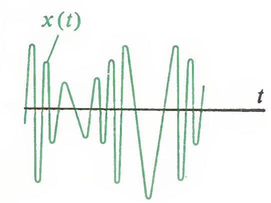
Then,

$$K(\tau)=\sigma_x^2\rho(\tau)\cos\omega_0\tau \tag{7.40}$$

The envelope and the initial phase. The form of the autocorrelation function in (7.40) implies that the individual realizations of a narrowband random process are quasi-harmonic waves

$$x(t)=U(t)\cos[\omega_0 t+\varphi(t)] \tag{7.41}$$

in which both the envelope $U(t)$ and the initial phase $\varphi(t)$ are random functions varying slowly (in scale of ω_0).



A typical realization of a narrowband random process

Let us write the realization defined in (7.41) as the sum of its in-phase and quadrature components (see Chap. 5):

$$x(t)=A(t)\cos\omega_0 t-B(t)\sin\omega_0 t \tag{7.42}$$

The two amplitudes, $A(t)$ and $B(t)$, are likewise slowly varying signals, and their variations grow progressively slower as the effective bandwidth $\Delta\omega_{\text{eff}}$ decreases in comparison with the centre frequency ω_0.

Consider a random process $Y(t)$ which is the conjugate of the original random process $X(t)$. Its realizations are found by the Hilbert transform:

$$y(t)=\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{x(\tau)\,d\tau}{t-\tau}$$

The assumption that the in-phase amplitude $A(t)$ and the quadrature amplitude $B(t)$ vary slowly enable us to write a fairly simple expression for a realization of the conjugate process by placing the slowly changing terms outside the Hilbert transform

$$y(t)=A(t)\sin\omega_0 t+B(t)\cos\omega_0 t \tag{7.43}$$

Hence follow the expressions for the instantaneous values of a realization of the envelope

$$U(t)=\sqrt{x^2+y^2}=\sqrt{A^2(t)+B^2(t)} \tag{7.44}$$

and of the initial phase

$$\varphi(t)=\arc\tan\frac{B(t)}{A(t)} \tag{7.45}$$

Statistical properties of the conjugate process. For a further analysis of the envelope and initial phase of a narrowband random process, we must establish the relation between the statistical characteristics of $X(t)$ and $Y(t)$.

To begin with, it is to be noted that if $\bar{x}=0$, then $\bar{y}=0$ as well.

Also, since $X(t)$ is a Gaussian process, and the Hilbert transformation is a linear integral transformation, then the conjugate process $Y(t)$ has the property of normality as well.

As will be recalled, if $S_x(\omega)$ is the spectrum of the realization $x(t)$, then the corresponding spectrum of the conjugate realization will be

$$S_y(\omega) = -jS_x(\omega)\,\mathrm{sgn}(\omega)$$

The magnitudes of $S_x(\omega)$ and $S_y(\omega)$ are the same, so the power spectra of $X(t)$ and $Y(t)$ are the same, too:

$$W_x(\omega) = W_y(\omega)$$

Hence we may conclude that the two processes have an identical autocorrelation function and that $Y(t)$ *is a stationary process:*

$$K_x(\tau) = K_y(\tau) = \int_0^{\infty} F_x(\omega)\cos\omega\tau\,d\omega$$

Finally, let us find the cross-correlation function

$$K_{xy}(\tau) = \overline{x(t)\,y(t+\tau)} = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{\overline{x(t)\,x(\xi)}}{t+\tau-\xi}\,d\xi$$

$$= \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{K_x(\xi-t)}{\tau-(\xi-t)}\,d\xi = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{K_x(\eta)}{\tau-\eta}\,d\eta$$

which is the same as the Hilbert transform of the autocorrelation function of the process $X(t)$. Similarly, it can be shown (this is left as an exercise for the reader) that

$$K_{yx}(\tau) = -K_{xy}(\tau)$$

Thus,

$$K_{xy}(\tau) = H[K_x(\tau)] = \int_0^{\infty} F_x(\omega)\sin\omega\tau\,d\omega \tag{7.46}$$

Interestingly, the function $K_{xy}(\tau)$ shows odd symmetry and vanishes at $\tau = 0$. Hence, the processes $X(t)$ and $Y(t)$ are statistically independent at the corresponding instants of time. Equation (7.46) can be re-cast in a more convenient form on making a change of variable, $\omega = \omega_0 + \Omega$. Then

$$K_{xy}(\tau) = \int_{-\omega_0}^{\infty} F_x(\omega_0+\Omega)\sin(\omega_0+\Omega)\tau\,d\Omega$$

$$= a(\tau)\sin\omega_0\tau + b(\tau)\cos\omega_0\tau \tag{7.47}$$

where the functions $a(\tau)$ and $b(\tau)$ are defined as in (7.38).

Correlation properties of the in-phase and quadrature amplitudes. We seek to find and analyse the statistical characteristics of the envelope $U(t)$ and of the initial phase $\varphi(t)$. For this, it is convenient to replace the realizations $x(t)$ and $y(t)$ with the slowly varying realizations $A(t)$ and $B(t)$ which, on the basis of Eqs. (7.42) and (7.43), can be defined as

$$\begin{aligned} A(t) &= x(t)\cos\omega_0 t + y(t)\sin\omega_0 t \\ B(t) &= -x(t)\sin\omega_0 t + y(t)\cos\omega_0 t \end{aligned} \tag{7.48}$$

The processes $A(t)$ and $B(t)$ are linearly related to the Gaussian processes $X(t)$ and $Y(t)$. Therefore, they, too, are Gaussian, and if $\bar{x} = \bar{y} = 0$, then $\bar{A} = \bar{B} = 0$, as well.

Let us take the first line in Eqs. (7.48) and find the autocorrelation function of the process $A(t)$. By simple trigonometry, we obtain

$$K_A(\tau) = \overline{[x(t)\cos\omega_0 t + y(t)\sin\omega_0 t]\,[x(t+\tau)\cos\omega_0(t+\tau)} \\ \overline{+\,y(t+\tau)\sin\omega_0(t+\tau)]}$$

$$= K_x(\tau)\cos\omega_0\tau + K_{xy}(\tau)\sin\omega_0\tau \tag{7.49}$$

Substituting for $K_x(\tau)$ and $K_{xy}(\tau)$ from (7.38) and (7.47) yields a very simple result:

$$K_A(\tau) = a(\tau) \tag{7.50}$$

Similarly, it is proved that

$$K_B(\tau) = a(\tau) \tag{7.51}$$

▲ **Work Problem 10**

and

$$K_{AB}(\tau) = b(\tau) \tag{7.52}$$

On setting $\tau = 0$ in (7.50) and (7.51), we obtain

$$\sigma_A^2 = \sigma_B^2 = \int_{-\infty}^{\infty} F(\omega)\,d\omega = \sigma_x^2 \tag{7.53}$$

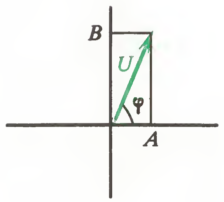
Thus, the variances of the in-phase and quadrature amplitudes are equal to the variance of the original narrowband process $X(t)$.

The joint probability density of the envelope and of the initial phase. The advantages of the procedure in which the narrowband random process is replaced with its in-phase and quadrature components become obvious when we are to find the bivariate probability density $p(U, \varphi)$. This characteristic, in turn, enables us to find the univariate probability densities of the envelope

$$p(U) = \int_0^{2\pi} p(U, \varphi)\,d\varphi \tag{7.54}$$

and of the initial phase

$$p(\varphi)=\int_0^\infty p(U,\ \varphi)\,dU \tag{7.55}$$



The instantaneous values of $A(t)$ and $B(t)$ form a two-dimensional Gaussian vector variable the two components of which are independent and have the same variance σ_x^2. Therefore, the bivariate probability density is

$$p(A,\ B)=p(A)\,p(B)=\frac{1}{2\pi\sigma_x^2}\exp[-(A^2+B^2)/2\sigma_x^2] \tag{7.56}$$

In order to obtain the probability density $p(U,\ \varphi)$, we should transform the random vector $\{A, B\}$ into a new random vector $\{U, \varphi\}$:

$$A=U\cos\varphi,\quad B=U\sin\varphi \tag{7.57}$$

The Jacobian of such a transformation (see Eq. (6.24)) is

$$D=\begin{vmatrix}\cos\varphi & -U\sin\varphi\\ \sin\varphi & U\cos\varphi\end{vmatrix}=U \tag{7.58}$$

Since in the new variables

$$A^2+B^2=U^2$$

it follows then that the expression for the sought-for bivariate probability density is

$$p(U,\ \varphi)=(U/2\pi\sigma_x^2)\exp(-U^2/2\sigma_x^2) \tag{7.59}$$

The univariate phase distribution. Taking advantage of Eqs. (7.55) and (7.59), let us find the probability density of the initial phase:

$$p(\varphi)=\frac{1}{2\pi}\int_0^\infty (U/\sigma_x^2)\exp(-U^2/2\sigma_x^2)\,dU$$

By a change of variable, $z=U/\sigma_x$, we get

$$p(\varphi)=\frac{1}{2\pi}\int_0^\infty z\exp(-z^2/2)\,dz=\frac{1}{2\pi} \tag{7.60}$$

Thus, the initial phase of a narrowband random process is *distributed uniformly* in the interval from 0 to 2π. Physically, this means that all values of the initial phase of individually taken realizations are equally significant.

The univariate envelope distribution. Since $p(U,\ \varphi)$ is not an explicit function of φ, then, by virtue of Eq. (7.54), the probability

density of the envelope is

$$p(U)=(U/\sigma_x^2)\exp(-U^2/2\sigma_x^2) \tag{7.61}$$

It is advisable to change back to the nondimensional variable $z=U/\sigma_x$, for which

$$p(z)=z\exp(-z^2/2) \tag{7.62}$$

The Rayleigh distribution

The distribution of the instantaneous values of the envelope of a narrowband random process as defined by Eq. (7.61) or (7.62) is known as the *Rayleigh distribution*. The applicable plot in Fig. 7.1 demonstrates that only some of the average values of the envelope (the orders of σ_x) are most probable. On the other hand, it is very unlikely that the envelope can take on values either very close to zero or substantially exceeding the rms level, σ_x, of the original process.

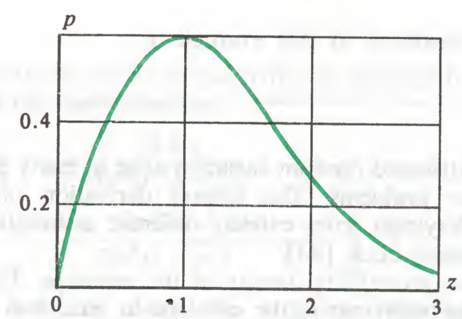


Fig. 7.1 Plot of the probability density for a Rayleigh-distributed random variable (with the dimensionless argument $z=U/\sigma_x$ laid off as abscissa)

By averaging on the basis of the distribution in (7.61), we can find the mean value of the envelope of a narrowband random process:

$$m_U=\bar{U}=\sqrt{\pi/2}\,\sigma_x=1.253\sigma_x \tag{7.63}$$

Solve Problem 11

and its variance:

$$\sigma_U^2=\bar{U^2}-\bar{U}^2=(2-\pi/2)\,\sigma_x^2=0.429\sigma_x^2 \tag{7.64}$$

The univariate probability density of the envelope enables us to solve many problems in the theory of narrow-band random processes, notably to find the probability of the envelope exceeding a definite specified level.

Example 7.5. *Let a narrowband normal process have a constant one-sided power spectrum $F_0 = 1.5 \times 10^{-3}\ \mathrm{V^2\ s}$ in the frequency band from $\omega_{\min} = 10^5\ \mathrm{s^{-1}}$ to $\omega_{\max} = 1.02 \times 10^5\ \mathrm{s^{-1}}$. Find the probability that the envelope of this process will exceed the level $U_0 = 5$ V.*

From the statement of the problem, the effective band-width of the process is

$$\Delta\omega_{\text{eff}} = 2 \times 10^3\ \mathrm{s^{-1}}$$

Therefore, the variance is

$$\sigma_x^2 = F_0 \Delta\omega_{\text{eff}} = 3\ \mathrm{V^2}$$

From the definition of the probability density, the sought-for quantity is

$$P(U \geqslant U_0) = \int_{U_0}^{\infty} p(U)\,dU = \exp(-U_0^2/2\sigma_x^2)$$

$$= \exp(-25/6) = 0.0155$$

It may be argued that the event considered in this example is a rather seldom occurrence.

Rayleigh-distributed random variables arise in many physical and communication problems. The elegant derivation of Eq. (7.61) obtained by Rayleigh from entirely different assumptions can be found in a classic book [43].

The bivariate probability density of the envelope. For a proper insight into the behaviour of the envelope in time, it is essential to have a more detailed information as compared with that obtainable from the Rayleigh distribution. For example, if we wish to calculate the autocorrelation function of the envelope, we need to know the bivariate probability density $p(U, U_\tau)$. (Here and elsewhere, $U_\tau = U(t+\tau)$.)

Let us take advantage of the fact that the in-phase and quadrature amplitudes of a narrowband normal process are low-frequency Gaussian signals having the same autocorrelation function

$$K_A(\tau) = K_B(\tau) = \sigma_x^2 \rho(\tau)$$

and the following bivariate densities [see (6.28)]:

$$p(A, A_\tau) = \frac{1}{2\pi\sigma_x^2\sqrt{1-\rho^2}} \exp\left[-\frac{A^2 + A_\tau^2 - 2\rho A A_\tau}{2\sigma_x^2(1-\rho^2)}\right]$$

$$p(B, B_\tau) = \frac{1}{2\pi\sigma_x^2\sqrt{1-\rho^2}} \exp\left[-\frac{B^2 + B_\tau^2 - 2\rho B B_\tau}{2\sigma_x^2(1-\rho^2)}\right]$$

If the power spectrum is symmetrical about the centre frequency ω_0, such that $b(\tau) = 0$, the processes $A(t)$ and $B(t)$ are statistically independent, and so their joint quadrivariate probability density is

$$p(A, A_\tau, B, B_\tau) = p(A, A_\tau)\,p(B, B_\tau) \tag{7.65}$$

Let us change from the in-phase and quadrature amplitudes to the envelope and the phase observed at different instants of time:

$$\begin{aligned} A &= U\cos\varphi; \quad A_\tau = U_\tau \cos\varphi_\tau \\ B &= U\sin\varphi; \quad B_\tau = U_\tau \sin\varphi_\tau \end{aligned} \tag{7.66}$$

The Jacobian of the above transformation is

$$D = \begin{vmatrix} \cos\varphi & 0 & \sin\varphi & 0 \\ 0 & \cos\varphi_\tau & 0 & \sin\varphi_\tau \\ -U\sin\varphi & 0 & U\cos\varphi & 0 \\ 0 & -U_\tau \sin\varphi_\tau & 0 & U_\tau\cos\varphi_\tau \end{vmatrix} = UU_\tau \tag{7.67}$$

Using the above result, we can write the probability density (7.65) in terms of the new variables:

$$p(U, U_\tau, \varphi, \varphi_\tau) = \frac{UU_\tau}{4\pi^2\sigma_x^4(1-\rho^2)} \times \exp\left[-\frac{U^2 + U_\tau^2 - 2\rho\, UU_\tau \cos(\varphi - \varphi_\tau)}{2\sigma^2(1-\rho^2)}\right] \tag{7.68}$$

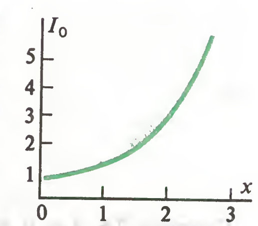
Now, in order to obtain the sought-for bivariate probability density of the envelope, we should integrate the right-hand side of Eq. (7.68) with respect to angular coordinates twice:

$$p(U, U_\tau) = \int_0^{2\pi} d\varphi \int_0^{2\pi} d\varphi_\tau\, p(U, U_\tau, \varphi, \varphi_\tau) \tag{7.69}$$

By applying a representation known from mathematics

$$\frac{1}{2\pi}\int_0^{2\pi} \exp(x\cos\varphi)\,d\varphi = I_0(x)$$

where I_0 is a modified Bessel function of order zero, from Eqs. (7.68) and (7.69) we finally get



The function $I_0(x)$

$$p(U, U_\tau) = \frac{UU_\tau}{\sigma_x^4(1-\rho^2)} \exp\left[-\frac{U^2 + U_\tau^2}{2\sigma_x^2(1-\rho^2)}\right] I_0\left[\frac{\rho UU_\tau}{\sigma_x^2(1-\rho^2)}\right] \tag{7.70}$$

The bivariate Rayleigh distribution

The probability density defined by Eq. (7.70) is sometimes called the *bivariate Rayleigh distribution*. Interestingly, if the time shift τ substantially exceeds the correlation time τ_{cor} associated with the function $\rho(\tau)$, then $\rho \to 0$ and, since $I_0(0)=1$, we have

$$p(U, U_\tau) \approx (U/\sigma_x^2)\exp(-U^2/2\sigma_x^2)(U_\tau/\sigma_x^2)\exp(-U_\tau^2/2\sigma_x^2) \qquad (7.71)$$

that is, $p(U, U_\tau)$ tends towards the product of two univariate Rayleigh densities.

The autocorrelation function of the envelope. By definition,

$$K_U(\tau) = \overline{UU_\tau} - \overline{U}^2 \qquad (7.72)$$

On the basis of Eq. (7.63), the square of the mean of the envelope is

$$\overline{U}^2 = (\pi/2)\sigma_x^2$$

Therefore, the task reduces to calculating the mean value of the product UU_τ:

$$\overline{UU_\tau} = \int_0^\infty dU \int_0^\infty dU_\tau p(U, U_\tau) UU_\tau \qquad (7.73)$$

The evaluation of the integral in (7.73) involves a good deal of computational work because the bivariate probability density (7.70) expands into an infinite series of Laguerre polynomials [15]. Omitting the intervening manipulations, the final result can be written as

$$K_U(\tau) = \frac{\pi}{2}\sigma_x^2\left\{\frac{\rho^2}{4} + \sum_{n=2}^{\infty}\frac{[(2n-3)!!]^2}{2^{2n}(n!)^2}\rho^{2n}\right\} \qquad (7.74)$$

On writing the autocorrelation function as

$$K_U(\tau) = \sigma_U^2 R_U(\tau)$$

we find from (7.74) the following expression for the correlation coefficient of the envelope:

Solve Problem 12

$$R_U(\tau) = 0.915\rho^2(\tau) + 0.058\rho^4(\tau) + \ldots \qquad (7.75)$$

In approximate calculations, it is legitimate to take the correlation coefficient of the envelope as being equal to the square of the envelope of the autocorrelation function of the narrowband signal.

Example 7.6. *Let it be known that the autocorrelation function (in* V^2) *of a random process is*

$$K_x(\tau) = 0.5\exp(-10^4|\tau|)\cos 2\pi \times 10^6\tau$$

Since the high-frequency component has a period of 10^{-6} s, and the amplitude coefficient changes in the meantime by a factor of only $\exp(-10^{-2}) = 0.99$, the random process in question may be taken as a real narrowband process with a centre frequency $f_0 = 1$ MHz. Retaining only the first term in the series (7.75) and replacing the coefficient 0.915 approximately with unity, the correlation coefficient of the envelope is found to be

$$R_U(\tau) \approx \exp(-2\times 10^4|\tau|)$$

The variance of the envelope is

$$\sigma_U^2 = 0.429\sigma_x^2 = 0.2145\ \mathrm{V}^2$$

Hence, the autocorrelation function of the envelope is

$$K_U(\tau) = 0.1963\exp(-2\times 10^4|\tau|)$$

The correlation time of the envelope

$$\tau_{cor} = \int_0^\infty R_U(\tau)\,d\tau = 0.5\times 10^{-4}\ \mathrm{s}$$

is equal to 50 periods of the harmonic wave of frequency f_0. Finally, the one-sided power spectrum of the envelope (see Example 7.1), whose dimension is V^2 s,

$$F_U(\omega) = \frac{2.73\times 10^3}{4\times 10^8 + \omega^2}$$

is concentrated in the low-frequency region.

The envelope of the sum of a harmonic signal and a narrowband Gaussian noise. In communications engineering, it is often important to know the statistical regularities of the signal observed at the output of a frequency-selective device, such as a tuned amplifier. Here, the Gaussian fluctuation noise at central frequency ω_0 equal to the resonance frequency of the system is present along with a deterministic signal $U_m \cos\omega_0 t$ of a known amplitude U_m. If the message transmitted over the communication system is embedded in the envelope of the output wave, as is the case with amplitude modulation, it is important to learn the extent of signal corruption due to noise.

The simplest problem is to find the univariate probability density for the envelope of the aggregate waveform. Assuming that the signal is

$$s(t) = U_m\cos\omega_0 t$$

and the fluctuation noise is

$$n(t) = A(t)\cos\omega_0 t - B(t)\sin\omega_0 t$$

we may write the following expression for a realization of the

aggregate process $x(t)$:

$$x(t)=s(t)+n(t)=[U_m+A(t)]\cos\omega_0 t-B(t)\sin\omega_0 t$$

This is a narrowband process, therefore its realization may be written in terms of the slowly changing envelope $U(t)$ and of the initial phase $\varphi(t)$:

$$x(t)=U(t)\cos[\omega_0(t)+\varphi(t)]$$

Obviously, there is a relation between the pairs $\{A, B\}$ and $\{U, \varphi\}$ such that

$$A=U\cos\varphi-U_m$$

$$B=U\sin\varphi \tag{7.76}$$

It is an easy matter to verify that the Jacobian D of the above transformation is equal to U. In consequence, since the bivariate probability density is

$$p(A,B)=\frac{1}{2\pi\sigma_x^2}\exp[-(A^2+B^2)/2\sigma_x^2]$$

then in the new variables

$$p(U,\ \varphi)=\frac{U}{2\pi\sigma_x^2}\exp\left(\frac{U^2+U_m^2-2UU_m\cos\varphi}{2\sigma_x^2}\right) \tag{7.77}$$

Now, in order to find the univariate probability density of the envelope, we should integrate the right-hand side of Eq. (7.77) with respect to the angular coordinate:

$$p(U)=\int_0^{2\pi}p(U,\ \varphi)\,d\varphi$$

As a result, we get

The Rice distribution

$$p(U)=(U/\sigma_x^2)\exp[-(U^2+U_m^2)/2\sigma_x^2]\,I_0(UU_m/\sigma_x^2) \tag{7.78}$$

Equation (7.78) expresses the distribution law known in communication theory as the *Rice distribution*. It is to be noted that for $U_m=0$, that is, in the absence of a deterministic signal, the Rice distribution goes over into the Rayleigh distribution automatically.

Figure 7.2 shows the probability density function of a Rice-distributed random variable for several values of the ratio $\alpha=U_m/\sigma_x$.

Interestingly, if the amplitude of the deterministic signal is substantially greater than the rms noise level, that is, if $U_m/\sigma_x\gg 1$, then for $U\gg U_m$ we may use the asymptotic representation for

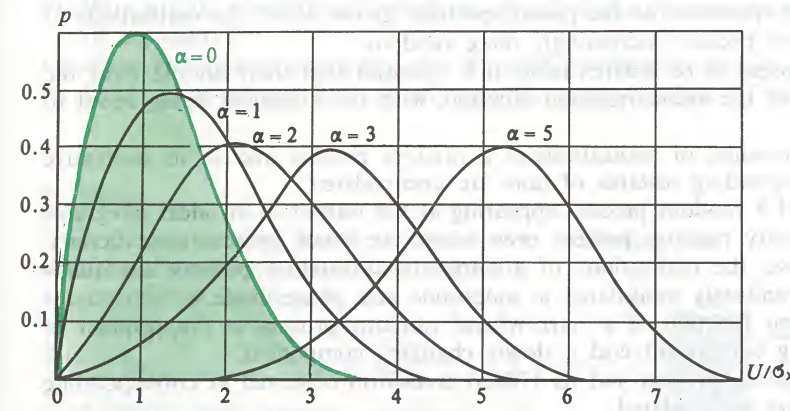


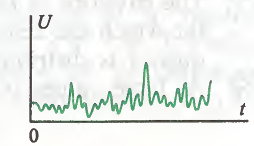
Fig. 7.2 Probability density of a Rice-distributed random variable

The curve in colour answers the Rayleigh distribution

modified Bessel functions with a large argument:

$$I_0(UU_m/\sigma_x^2)\approx\frac{\exp(UU_m/\sigma_x^2)}{\sqrt{2\pi(UU_m/\sigma_x^2)}}\approx\frac{\sigma_x\exp(UU_m/\sigma_x)}{\sqrt{2\pi U}}$$

A typical realization of the envelope of a narrowband noise

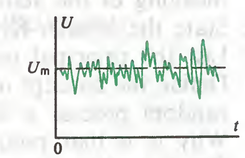


On substituting in (7.78), we get

$$p(U)\approx\frac{1}{\sqrt{2\pi}\sigma_x}\exp[-(U-U_m)^2/2\sigma_x^2] \tag{7.79}$$

That is, the envelope of the resultant signal is now an approximately Gaussian variable of variance σ_x^2 and of mean value U_m.

and a process which is the sum of a harmonic signal and narrowband noise



For practical purposes it is legitimate to assume that even at $U_m/\sigma_x=3$ the envelope of a narrowband signal is normalized. It is useful to recall that the Rayleigh-distributed envelope of pure noise has a variance of $0.429\sigma_x^2$. Thus, the increase in the amplitude of the deterministic harmonic signal more than doubles the variance of the envelope. Yet, the relative fluctuations of the envelope decrease. To demonstrate, for pure noise the ratio $\sigma_U/\overline{U}$, which can serve as a convenient numerical measure of fluctuations, is 0.523. In the case of a large deterministic signal, $\sigma_U/\overline{U}=\sigma_x/U_m$, tending to zero as U_m increases.

Note variations in the mean and the variance due to the harmonic signal

Summary

- The random spectrum of the individual realizations of a stationary random process is delta-correlated and has mean value zero at all frequencies.
- The Fourier transform of the autocorrelation function of a stationary random process

is called the power spectrum. As the power spectrum grows wider, the realizations of the random process become increasingly more random.

- For a random process to be differentiable, it is essential that there should exist the second derivative of the autocorrelation function, with the argument being equal to zero.
- The instantaneous values of realizations of a random process and of its derivative observed at corresponding instants of time are uncorrelated.
- The realizations of a random process appearing at the output of an ideal integrator form a nonstationary random process even when the input process is stationary.
- In the general case, the realizations of a narrowband random process are quasi-harmonic waves randomly modulated in amplitude and phase angle.
- The autocorrelation function of a narrowband random process is the product of a rapidly changing component and a slowly changing component.
- A narrowband random process and its Hilbert transform observed at corresponding instants of time are uncorrelated.
- The envelope of a narrowband Gaussian random process is Rayleigh-distributed; the initial phase of the process is uniformly distributed.
- The autocorrelation function of the envelope of a narrowband process is approximately equal to the square of the envelope of the autocorrelation function of the process itself.
- The envelope of the sum of a harmonic signal and of a narrowband Gaussian noise for which the centre frequency of the power spectrum is the same as the signal frequency is distributed by the Rice law.
- At large values of the signal-to-noise ratio, the envelope is normalized.

Review Questions

1. A random process is studied within the framework of correlation theory. What is the meaning of the statement?
2. State the Wiener-Khinchin (W-K) theorem.
3. List the principal properties of the power spectrum of a stationary random process.
4. Define the concept of one-sided power spectrum. If the power spectrum of a stationary random process is known, how can you find its variance?
5. Why is it that random processes resembling white noise are called delta-correlated? What are the basic properties of white noise? In what cases may a real random process be approximated with white noise?
6. Define the concept of convergence in probability.
7. List the salient properties of nondifferentiable random processes.
8. How does one find the variance, the autocorrelation function and the power spectrum of the derivative of a stationary random process?
9. Define the notion of an upslope (positive-going) crossing of a random process.
10. What is the quasi-frequency of a stationary random process?
11. Draw an approximate plot for a realization of a narrowband random process.
12. List the basic properties of the in-phase and quadrature amplitudes of a narrowband random process.

13. Why is it that the Hilbert transform of a Gaussian random process retains the property of normality?
14. Draw the typical waveform of a Rayleigh-distributed random signal.
15. Draw plots of the Rice distribution for several values of the signal-to-noise ratio.

Problems

1. Calculate the power spectrum of a stationary random process for which the autocorrelation function is

$$K(\tau)=\begin{cases}\sigma^2(1-|\tau|/t_0), & |\tau|\leqslant t_0\\ 0, & |\tau|>t_0\end{cases}$$

2. Find the power spectrum of a stationary random process for which the autocorrelation function is

$$K(\tau)=\sigma^2\exp(-\beta^2\tau^2/2)\cos\omega_0\tau$$

3. The one-sided power spectrum of a stationary random process $X(t)$ is defined by

$$F_X(f)=A(f/f_0)\exp(-f/f_0)$$

where A and f_0 are constants. Determine the autocorrelation function of the process.

4. A stationary random process has an effective bandwidth of 1.5 MHz. The maximum value of the one-sided power spectrum is $0.3\times10^{-12}\ \mathrm{V^2\,Hz^{-1}}$. Determine the variance of the process.

5. Find the correlation time for a stationary random process whose autocorrelation function has the form

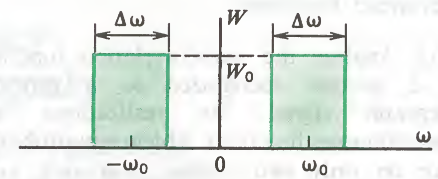
$$K(\tau)=\sigma^2\exp(-\alpha|\tau|)$$

6. Assuming that the air temperature over a month is a realization of a stationary random process, estimate the correlation time.

7. Find the variance for the derivative of a random process whose autocorrelation function has the form

$$K(\tau)=\sigma^2(1+\alpha|\tau|)\exp(-\alpha|\tau|)$$

8. A stationary random process has a power spectrum represented by the following plot:



Prove that this random process is differentiable, and find the variance of its derivative.

9. The quasi-frequency of a random process of variance 8 V² is 0.5 MHz. Determine the variance of its derivative.

10. A stationary random process $X(t)$ has the power spectrum

$$F_X(\omega)=\begin{cases}F_0 & \text{for } \omega_1<\omega<\omega_2\\ 0 & \text{for } \omega<\omega_1,\ \omega>\omega_2\end{cases}$$

The realizations of the process are defined by

$$x(t)=A(t)\cos\omega_0 t-B(t)\sin\omega_0 t$$

where

$$\omega_0=(\omega_2-\omega_1)/2$$

Find its autocorrelation functions $K_A(\tau)$ and $K_B(\tau)$ and its cross-correlation function $K_{AB}(\tau)$.

11. Find the mean and the variance of the envelope of a narrowband normal random process for which the autocorrelation function is

$$K_X(\tau)=25\exp(-4\times10^6\tau^2)\cos10^9\tau$$

12. A narrowband normal random process $X(t)$ has the autocorrelation

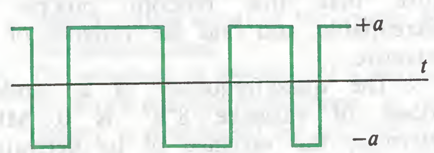
function

$$K_x(\tau) = \sigma^2 \exp(-\beta\tau^2/2)\cos\omega_0\tau$$

Find the autocorrelation function and the power spectrum for the envelope of the process.

Advanced Problems

13. Analyse the autocorrelation function of a process recognized as a random telegraph signal. Its realizations are discontinuous functions which equiprobably take on only two values: $+a$ and $-a$:



At random instants of time, each realization changes its sign abruptly. The probability that over a time T the sign will be changed n times obeys the Poisson law:

$$P_T(n) = \frac{(\lambda T)^n}{n!}\exp(-\lambda T)$$

where λ is the parameter defining the time rate of change of the process.

14. A narrowband normal random process $X(t)$ has the autocorrelation function

$$K_x(\tau) = \sigma_x^2 \rho(\tau)\cos\omega_0\tau$$

Prove that the square of its envelope has the autocorrelation function

$$K_{U^2}(\tau) = 4\sigma_x^2\rho^2(\tau)$$

15. A narrowband normal random process has the autocorrelation function defined in Problem 14. Find the univariate probability density for the instantaneous frequency of the process.

2. Circuits

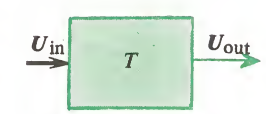
Chapter 8

Response of Linear Stationary Systems to Deterministic Signals

Part One of this book has dealt with the basics of signal theory. It has been stressed that signals can be the objects of a theoretical study only after we introduce suitable mathematical models. This is also true of the structures used to process, transform or transmit signals. These structures are extremely diverse in both their internal arrangement and in their external characteristics. So that they can be compared and classified, we should above all formulate a number of basic notions.

8.1 Physical Systems and Their Mathematical Models

However different communication circuits may be, each will always have an *input* to receive the original signals, and an *output* to deliver the transformed signals for further use. As a rule, this is illustrated by a block diagram of the "black box" type.



System operators. In the simplest case, both the input signal $u_{in}(t)$ and the output signal $u_{out}(t)$, frequently called the *response* of a system, are individual functions of time. In the more general case, the input signal is represented as an m-dimensional or vector variable

$$\breve{U}_{in}(t) = \{u_{in,1}(t), u_{in,2}(t), \ldots, u_{in,m}(t)\}$$

and the output signal as an n-dimensional or vector variable:

$$\breve{U}_{out}(t) = \{u_{out,1}(t), u_{out,2}(t), \ldots, u_{out,n}(t)\}$$

A system operator

The relation between $\breve{U}_{in}$ and $\breve{U}_{out}$ can be specified by means of a *system operator* T whose action on $\breve{U}_{in}$ results in $\breve{U}_{out}$

$$\breve{U}_{out}(t) = T\breve{U}_{in}(t) \tag{8.1}$$

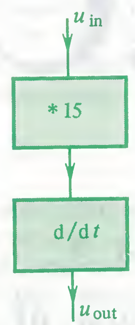
Example 8.1. *Suppose that a system transforms a one-dimensional signal such that*

$u_{out}(t) = 15\,du_{in}(t)/dt$

In this case the system operator may be written as

$T \equiv 15\,d/dt$

From the above expression we can identify the system as consisting of a scaler (an ideal amplifier) and a differentiator, as shown in the accompanying block diagram.



To make the problem completely defined, it is also necessary to specify the domain D_{in} in some functional (for example, Hilbert) space, which is called the *domain of eligible input signals*. In the simplest case it is specified by stating the kind of eligible input signals which may be continuous or discrete, deterministic or random. Similarly we define the *domain of eligible output signals*, D_{out}.

A mathematical model of a physical system

The combination of an applicable system operator T and the two eligible signal domains, D_{in} and D_{out}, will be further referred as a *mathematical model* of the physical system involved, and communication systems will be classed on the basis of the most significant properties of their mathematical models.

More specifically, this Chapter will be solely concerned with systems acted upon by analog signals defined by both continuous and discontinuous functions of time. The transformation of discrete (and, especially, digital) signals by linear systems will be examined in Chap. 15.

Stationary and nonstationary systems. A system is said to be *stationary*, if its response is independent of the time at which it accepts the input signal. If T is the operator of a stationary system, then from the equality

$$\vec{U}_{out}(t) = T\vec{U}_{in}(t) \tag{8.2}$$

it follows that

$$\vec{U}_{out}(t + t_0) = T\vec{U}_{in}(t \pm t_0) \tag{8.3}$$

for any value of t_0. Also, systems are called stationary if their parameters remain constant in time, or time-invariant.

Otherwise a system is called *nonstationary* (a time-variant or parametric system).

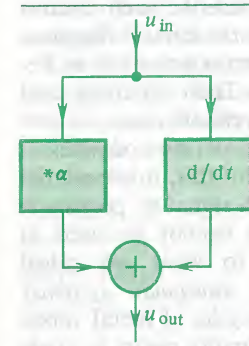
The two classes of systems are widely used in telecommunications and will be the subject of the subsequent discussion. It should be pointed out, however, that a theoretical study of nonstationary systems is ordinarily a far more complicated task.

Linear and nonlinear systems. A most important principle underlying the classification of communication systems is that various systems generally respond differently to a sum of input signals. If the system operator is such that

The principle of superposition

$$T(\vec{U}_{in,1} + \vec{U}_{in,2}) = T\vec{U}_{in,1} + T\vec{U}_{in,2}$$
$$T(\alpha\vec{U}_{in}) = \alpha T\vec{U}_{in} \tag{8.4}$$

where α is an arbitrary number, then the system is *linear*. The conditions in (8.4) define the fundamental *superposition principle.*



Example 8.2. *A system transforms the input signal in such a way that*

$$u_{out}(t) = \left(\frac{d}{dt} + \alpha\right)u_{in}(t)$$

It can be verified by inspection that the conditions defined in (8.4) are satisfied. Thus, the system in question is linear.

Example 8.3. *Suppose that a system is an ideal squarer operating in accordance with the algorithm*

$u_{out}(t) = u_{in}^2(t)$

By applying a sum of two input signals, $u_{in,1} + u_{in,2}$, we obtain the following response:

$u_{out}(t) = u_{in,1}^2 + 2u_{in,1}u_{in,2} + u_{in,2}^2$

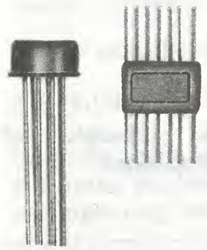
The presence of the cross-product $2u_{in,1}u_{in,2}$ is an indication that the system is nonlinear.

Strictly speaking, all physical systems dealt with in telecommunications are to a varying degree nonlinear. However, there is a class of systems for which linear models yield a fairly accurate description. For example, we can practically always ignore the nonlinear behaviour of ordinary resistors, capacitors and some inductors. As will be shown later, linear systems are remarkable in that for them we can, at least theoretically, solve any problem involving the transformation of an input signal.

Nonlinear circuits usually contain diodes and transistors. Their nonlinearity manifests itself in that their current-voltage characteristics strongly differ from linear relations.

Although the theory of nonlinear systems is, as a rule, very complicated, and far from any results can be deduced analytically, it is with the aid of nonlinear elements that the most important transformations are performed on communication signals. A theory of the simplest nonlinear systems is set forth in Chap. 11.

Lumped-constant and distributed-constant systems. Another principle for classifying the elements, devices and systems used in telecommunications is by comparing the physical size of a system with the wavelength applied to its input. If the characteristic dimension of a system (say, the maximum length of the wires in an electric circuit) is substantially smaller than the wavelength used, we have a *lumped-constant system.*



ICs are examples of lumped-constant systems

In a lumped-constant electric circuit, it is always possible to single out physical regions within which the energy of the electric field, as in capacitors, or the energy of a magnetic field, as in inductors, is concentrated (or lumped). The properties of lumped-constant circuits are independent of the configuration of connecting wires, therefore it is customary to describe such circuits by use of their abstract models called *schematic circuit diagrams.*

In telecommunications, lumped-constant circuits are used at frequencies up to several hundred megahertz. Their analysis and synthesis are carried out by invoking the Kirchhoff laws.

A waveguide is an example of a distributed-constant system

At microwave frequencies, the physical size of most devices is comparable with the wavelength used. Because of this, it is essential to consider the effect of the finite time it takes a signal to propagate around the system. Ordinary electric circuits cannot be used at microwave frequencies, and they give way to what are called *distributed-constant systems* (also known as *waveguide systems*). Instead of connecting wires, use is made of lengths of metal tubes, called waveguides, and instead of *LC* tuned circuits resort is made to their distributed-constant counterparts known as *cavity resonators* or *resonant cavities.* As a rule, the theory, analysis and synthesis of distributed-constant systems are fairly complex and are the subject-matter of separate disciplines.

To conclude our brief overview of system classification, we shall concentrate on lumped-constant linear stationary systems – the simplest type of all. The most typical examples are linear electric circuits whose properties must be known to the student from the previous courses. To them we may add, however, any linear physical devices and systems whose properties remain inchanged in time and whose size is small as compared with the wavelength used.

8.2 The Impulse, Step and Frequency Responses of Linear Stationary Systems

Linear systems have a remarkable property – they obey the superposition principle. This offers a direct approach to a systematic analysis of system responses to a variety of input signals, also called *driving* or *forcing functions*, or *excitations.* Using the dynamic representation of signals set forth in Chap. 1, we can resolve input signals into sums of elementary pulses. If, now, we find, in one way or another, the response of a system to any

elementary excitations, the final step in solving the problem will be simply to combine the individual responses.

This form of analysis is based on the representation of signal and system properties in the time domain. It is equally possible and, sometimes, even more convenient to use their analysis in the frequency domain when signals are specified by giving their Fourier series expansion or integrals. Then the system properties can be represented by their frequency characteristics which define how the elementary harmonic signals are transformed.

The impulse response. As will be recalled, the dynamic representation of signals uses any one of two elementary excitations. These are the *unit step input* and the *unit impulse input* or *delta function.* Let us first turn to the integral representation of a signal by means of the delta funcion.

Let a linear stationary system be described by its operator T. For simplicity we assume that the input and output signals are one-dimensional. By definition, the *impulse response* of a system is the function $h(t)$ which is the system response to a forcing function of the form $\delta(t)$. This implies that $h(t)$ satisfies the following equation:

In mathematics, the impulse response is known as the Green function of the operator in question

$$h(t) = T\delta(t) \tag{8.5}$$

Since the system is stationary, a similar equation will hold when the input excitation is translated in time through an arbitrary interval t_0:

$$h(t-t_0) = T\delta(t-t_0) \tag{8.6}$$

It must be clearly realized that the impulse response as well as the forcing delta-function are outcomes of a reasonable idealization. From a physical point of view, the impulse response approximates the response of a system to a unit-area impulse forcing function of an arbitrary waveform, provided that the duration of the input signal is negligible in comparison with the system's transient time, that is, the time it takes for the system to reach a steady state.

The physical significance of the impulse response

The Duhamel superposition integral. An extremely important role that the impulse response plays in the theory of linear stationary systems stems from the fact that knowledge of $h(t)$ enables us to solve formally any problem concerned with the passage of a deterministic signal through such a system. Indeed, as has been demonstrated in Chap. 1, the input signal can always be represented as

$$u_{\text{in}}(t) = \int_{-\infty}^{\infty} u_{\text{in}}(\tau)\,\delta(t-\tau)\,d\tau$$

The respective response is then

$$u_{\text{out}}(t) = Tu_{\text{in}}(t) = T\int_{-\infty}^{\infty} u_{\text{in}}(\tau)\,\delta(t-\tau)\,d\tau \tag{8.7}$$

Now we recall that an integral is the limit of a sum, so the linear operator T may, on the basis of the superposition principle, be included under the integral. Also, the operator T "operates" only on the terms dependent on the current time t, but not on the integration variable τ. Therefore, from Eq. (8.7) it follows that

$$u_{\text{out}}(t) = \int_{-\infty}^{\infty} u_{\text{in}}(\tau)\,T\delta(t-\tau)\,d\tau$$

or, finally

$$u_{\text{out}}(t) = \int_{-\infty}^{\infty} u_{\text{in}}(\tau)\,h(t-\tau)\,d\tau \tag{8.8}$$

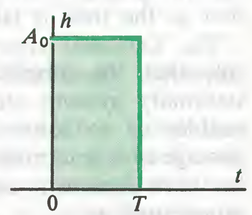
The result in Eq. (8.8), of fundamental significance in the theory of linear systems, is called the *Duhamel superposition integral.* From inspection of Eq. (8.8) it is seen that the output signal from a linear stationary system is the convolution of two functions – the input signal (excitation) and the impulse response. Obviously, we may re-write Eq. (8.8) as

▲ **Work Problem 1**

$$u_{\text{out}}(t) = \int_{-\infty}^{\infty} u_{\text{in}}(t-\tau)\,h(\tau)\,d\tau \tag{8.9}$$

A major advantage of the analysis based on the Duhamel superposition integral is that, once the impulse response $h(t)$ is found, we can reduce the subsequent computational steps to completely formalized operations. Even if the integrals in (8.8) and (8.9) cannot be found analytically, it is always possible to analyse them numerically on a computer.

Example 8.4. *Suppose that a linear stationary system, whose internal arrangement is immaterial, has an impulse response which is a rectangular video pulse of finite duration T and of amplitude A_0, occurring at $t = 0$:*



$$h(t) = \begin{cases} 0, & t < 0 \\ A_0, & 0 < t < T \\ 0, & t > T \end{cases}$$

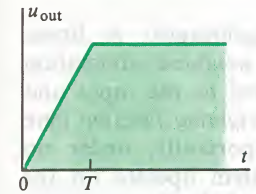
Find the response of the system, if the excitation is the step input (to

be defined later):

$$u_{\text{in}}(t) = U_0\,\sigma(t)$$

In using the Duhamel superposition integral in (8.8), we note that the output signal varies according as the time t exceeds or not the duration of the impulse response. For $0 < t < T$, we have

$$u_{\text{out}}(t) = A_0 U_0 \int_0^t d\tau = A_0 U_0 t$$



On the other hand, if $t > T$, then for $\tau > T$ the impulse function $h(t-\tau)$ vanishes, and so

$$u_{\text{out}}(t) = A_0 U_0 \int_0^T d\tau = A_0 U_0 T$$

The response thus found is plotted as a piecewise-linear function.

Generalization to the multidimensional (vector) case. It has been assumed that both the forcing function (excitation) and the output signal (response) are one-dimensional. In the more general case of a system with m inputs and n outputs, we need to introduce the partial impulse responses $h_{ij}(t)$ $(i = 1, 2, \ldots, n;\ j = 1, 2, \ldots, m)$, each of which represents the response at the ith output to the delta-function excitation which is applied to the jth input. The set of functions $h_{ij}(t)$ forms the impulse-response matrix

$$h(t) = \begin{bmatrix} h_{11} & h_{12} & \ldots & h_{1m} \\ \vdots & \vdots & & \vdots \\ h_{n1} & h_{n2} & \ldots & h_{nm} \end{bmatrix} \tag{8.10}$$

In the multidimensional (vector) case, the Duhamel superposition integral takes the form

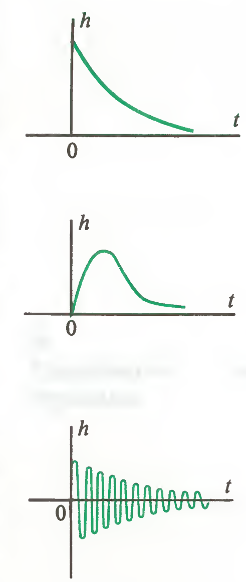
$$\dot{U}_{\text{out}}(t) = \int_{-\infty}^{\infty} h(t-\tau)\,\dot{U}_{\text{in}}(\tau)\,d\tau \tag{8.11}$$

where $\dot{U}_{\text{in}}$ is an m-dimensional vector, and $\dot{U}_{\text{out}}$ is an n-dimensional vector.

The condition for physical realizability. Whatever the actual form of the impulse response of a physically realizable system, the following important principle must always be satisfied: *A physical system cannot give an output in advance of the input signal.*

This principle imposes a very simple constraint on the form of feasible impulse responses:

$$h(t) = 0 \ \text{for}\ t < 0 \tag{8.12}$$



Examples of impulse responses for realizable systems

This condition is met, for example, by the impulse response of the system considered in Example 8.4.

It is easy to see that for a physically realizable system the upper limit in the Duhamel superposition integral must be replaced with the current value of time:

$$u_{out}(t) = \int_{-\infty}^{t} u_{in}(\tau) h(t-\tau)\,d\tau \tag{8.13}$$

Equation (8.11) has a clear physical significance: A linear stationary system performs the operation of weighted summation over all instantaneous excitation values applied to the input and existing in the "past" for $-\infty < \tau < t$. The *weighting function* here is the impulse response of the system. Importantly, under no circumstances can a physically realizable system operate on the information embodied in the "future" values of the input.

The step response. Let the forcing function (excitation) acting at the input of a linear stationary system be a *unit step input* represented by the Heaviside unit function $\sigma(t)$. The response of the system to this form of excitation

$$g(t) = T\sigma(t) \tag{8.14}$$

is called the *step response* of the system. Because the system in question is stationary, the step response is invariant under the time translation:

$$g(t-t_0) = T\sigma(t-t_0)$$

What has earlier been said about the realizability of a system fully applies to a system driven by a unit step input as well as by a unit impulse. Therefore, the step response of a realizable system is nonzero only for $t > 0$, whereas

$$g(t) = 0 \quad \text{for} \quad t < 0 \tag{8.15}$$

There is a close relation between the impulse response and the step response. To demonstrate, since $\delta(t) = d\sigma/dt$, then, by virtue of (8.5),

$$h(t) = T\left[\frac{d}{dt}\sigma(t)\right]$$

The differentiation operator d/dt and the linear stationary operator T may be interchanged, so

$$h(t) = \frac{d}{dt} T\sigma(t) = \frac{dg}{dt} \tag{8.16}$$

or

$$g(t) = \int_{-\infty}^{t} h(\xi)\,d\xi \tag{8.17}$$

Example 8.5. *Find the step response of the linear system from Example 8.4.*

Since here

$$h(t) = A_0\sigma(t) - A_0\sigma(t-T)$$

then, on the basis of Eq. (8.17),

$$g(t) = \begin{cases} 0 & \text{for } t < 0 \\ A_0 t & \text{for } 0 < t < T \\ A_0 T & \text{for } t > T \end{cases}$$

Thus, the system being at rest at $t < 0$ will change to a new stationary state in response to a unit step input during the time equal to the duration of the impulse response.

Taking advantage of Eq. (1.4) used in the dynamic representation of signals and proceeding along the same lines as in deriving Eq. (8.8), we obtain one more form for the Duhamel superposition integral:

$$u_{out}(t) = u_{in}(0)\,g(t) + \int_{0}^{t} (du_{in}/d\tau)\,g(t-\tau)\,d\tau \tag{8.18}$$

The frequency response. In mathematical analysis of systems, it is important to find such input signals (excitations) which, on being operated upon by a system, remain unchanged to within a certain numerical factor. If the equality

$$u_{out}(t) = Tu_{in}(t) = \lambda u_{in}(t) \tag{8.19}$$

Eigenfunction and eigenvalue

is satisfied, then $u_{in}(t)$ is called the *eigenfunction* of the operator T, and the number λ (complex in the general case) is called its *eigenvalue*.

Let us show that the complex signal $u_{in}(t) = \exp(j\omega t)$ is the eigenfunction of a linear stationary system for any ω. To this end, we take advantage of the Duhamel superposition integral as defined in (8.9) and obtain

$$u_{out}(t) = \int_{-\infty}^{\infty} \exp[j\omega(t-\tau)]\,h(\tau)\,d\tau$$

$$= \left[\int_{-\infty}^{\infty} h(\tau)\exp(-j\omega\tau)\,d\tau\right]\exp(j\omega t) \tag{8.20}$$

Hence, the eigenvalue is the complex number

The frequency response of a system

$$K(j\omega) = \int_{-\infty}^{\infty} h(t) \exp(-j\omega t)\, dt \tag{8.21}$$

called the *frequency response* of a system*.

Equation (8.31) establishes a fundamental fact: *The frequency response and the impulse response of a linear stationary system are related by a Fourier transform pair.* Therefore, if we know $K(j\omega)$, we can always determine the impulse response of the system

$$h(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} K(j\omega) \exp(j\omega t)\, d\omega \tag{8.22}$$

Analysis in the time domain and the frequency domain

This has led us to a crucial point in the theory we are concerned with: Any linear stationary system may be analysed either in the time domain by use of its impulse and step responses, or in the frequency domain on the basis of its frequency response. Each is as good as the other one, and the choice is a matter of convenience in obtaining initial specifications and calculations.

In conclusion, it should be noted that the frequency properties of a linear system having m inputs and n outputs can be described by the frequency response matrix

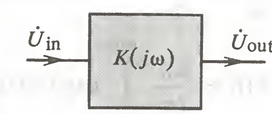
$$\underline{K}(j\omega) = \begin{bmatrix} K_{11} & K_{12} & \dots & K_{1m} \\ \vdots & \vdots & \vdots & \vdots \\ K_{n1} & K_{n2} & \dots & K_{nm} \end{bmatrix} \tag{8.23}$$

The relation between the matrices $\underline{h}$ and $\underline{K}$ is similar to that defined by Eqs. (8.21) and (8.22).

The amplitude (or magnitude) response and the phase response. The frequency response $K(j\omega)$ is simple to interpret: If the

* The alternative terms are the *complex response* or the *complex transfer function* (J. Brown and E. V. D. Glazier, *Telecommunications*. London: Chapman and Hall, 1974), the *system* or *response function* [20], the *transfer response* (*Handbook of Automation, Computation and Control.* Ed. Eu. M. Grabbe, S. Ramo, and D. E. Wooldridge. New York: John Wiley and Sons, Inc., 1958), etc.–*Translator's note.*

excitation is a harmonic signal of known frequency ω and of complex amplitude $\dot{U}_{in}$, then the complex amplitude of the output signal will be



$$\dot{U}_{out} = K(j\omega)\, \dot{U}_{in} \tag{8.24}$$

Frequently, especially in engineering calculations, the frequency response is written in the exponential form:

$$K(j\omega) = |K(j\omega)| \exp j\varphi_k(\omega) \tag{8.25}$$

The two real functions involved have each a special name. Thus, $|K(j\omega)|$ is called the *amplitude* (or *magnitude*) *response*, and $\varphi_k(\omega)$ is called the *phase response* of the system. There is a variety of instruments designed to measure and record the two functions of various electric and electronic circuits and devices in any frequency range.

The constraints imposed on the frequency response. Not just any function $K(j\omega)$ may be the frequency response of a realizable system. The simplest constraint arises because the impulse response $h(t)$ must be real. By virtue of the properties of Fourier transforms (see Chap. 2) this means that

$$K(j\omega) = K^*(-j\omega) \tag{8.26}$$

The properties of the amplitude and phase response

As follows from Eq. (8.26), the modulus of the frequency response (or the amplitude response) is an even function of frequency, whereas the phase angle (that is, the phase response) is an odd function of frequency.

The Paley-Wiener criterion

It is far more difficult to say what the frequency response must be like for the realizability condition, (8.12) or (8.15), to be satisfied. The answer is given by the *Paley-Wiener criterion* which we will give without proof. The Paley-Wiener criterion states that a necessary and sufficient condition for an amplitude response to be realizable is that

$$\int_{-\infty}^{\infty} \frac{|\log|K(j\omega)||}{1+\omega^2}\, d\omega < +\infty \tag{8.27}$$

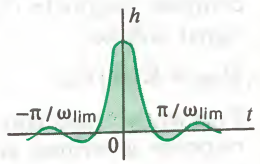
Let us consider a specific example which illustrates the relation between $K(j\omega)$ and $h(t)$.

Example 8.6. *Suppose that a linear stationary system behaves like an ideal low-pass filter. This implies that its frequency response is specified by the following set of equalities:*

$$K(j\omega) = \begin{cases} 0, & \omega < -\omega_{lim} \\ K_0, & -\omega_{lim} < \omega < \omega_{lim} \\ 0, & \omega > \omega_{lim} \end{cases}$$

On the basis of Eq. (8.20), the impulse response of such a filter is

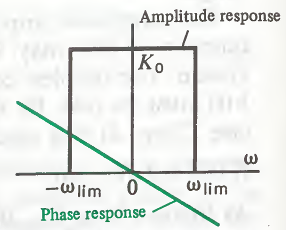
$$h(t)=\frac{K_0}{2\pi}\int_{-\omega_{\lim}}^{\omega_{\lim}}\exp(j\omega t)\,d\omega=\frac{K_0\omega_{\lim}}{\pi}\,\frac{\sin\omega_{\lim}t}{\omega_{\lim}t} \qquad (8.28)$$



The symmetry of the function about $t=0$ is an indication that an ideal low-pass filter is not physically realizable. In fact, the same conclusion could be drawn directly on the basis of the Paley-Wiener criterion. This is because the integral in (8.27) turns out to be divergent for any system whose amplitude response function vanishes within some finite interval on the frequency axis.

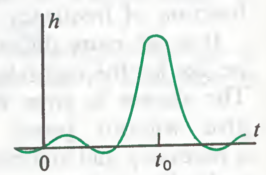
Although an ideal low-pass filter is not physically realizable, this model is successfully used to approximately describe the properties of filters, assuming that the function $K(j\omega)$ contains a phase factor which is a linear function of frequency:

$$K(j\omega)=\begin{cases}0, & \omega<-\omega_{\lim}\\ K_0\exp(-j\omega t_0), & -\omega_{\lim}<\omega<\omega_{\lim}\\ 0, & \omega>\omega_{\lim}\end{cases}$$



As can be readily verified, here

$$h(t)=\frac{K_0\omega_{\lim}}{\pi}\,\frac{\sin\omega_{\lim}(t-t_0)}{\omega_{\lim}(t-t_0)} \qquad (8.29)$$



The parameter t_0, equal to the slope of the phase response, defines the time delay for the time shift of the maximum of $h(t)$. Obviously, the accuracy with which this model can represent the properties of a realizable system improves as t_0 increases in value.

8.3 Linear Dynamic Systems

The heading refers to linear systems which have the following property: the output signal is determined not only by the value of the input signal at any particular instant of time at present, but also by the entire "past history" of the input process. In other words, a linear dynamic system has "memory" which governs the manner in which the input signal is transformed.

Systems described by differential equations. Of the many possible dynamic systems, those of primary concern to communication theory are systems which can be described by differential equations. In the most general case, these are systems for which the input-out-

put relations are established by the following differential equation:

$$a_n\frac{d^n u_{\text{out}}}{dt^n}+a_{n-1}\frac{d^{n-1}u_{\text{out}}}{dt^{n-1}}+\ldots+a_1\frac{du_{\text{out}}}{dt}+a_0u_{\text{out}}$$
$$=b_m\frac{d^m u_{\text{in}}}{dt^m}+b_{m-1}\frac{d^{m-1}u_{\text{in}}}{dt^{m-1}}+\ldots+b_1\frac{du_{\text{in}}}{dt}+b_0u_{\text{in}} \qquad (8.30)$$

Precisely this form of dynamic relation exists between the instantaneous values of the input and output signals in a lumped-constant electric circuit. If the circuit is linear and stationary, then all coefficients $a_0, a_1, \ldots, a_n$ and $b_0, b_1, \ldots, b_n$ are constant real numbers.

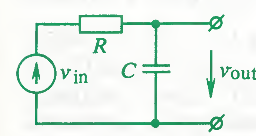
Suppose that the input signal $u_{\text{in}}(t)$ is fixed. Then the right-hand side of Eq. (8.30), which can arbitrarily be designated as $f(t)$, is a known function. Then the analysis of the system's behaviour reduces to the well-known mathematical problem of solving an nth-order linear differential equation with constant coefficients:

$$a_n\frac{d^n u_{\text{out}}}{dt^n}+a_{n-1}\frac{d^{n-1}u_{\text{out}}}{dt^{n-1}}+\ldots a_0u_{\text{out}}=f(t) \qquad (8.31)$$

The order n of the equation is called the *order of the dynamic system* involved.

The order of a dynamic system

Consider several examples of dynamic systems and the corresponding differential equations.



First-order systems are also called pure-delay (or lag) elements

Example 8.7. *Consider an RC-network in the form of an L-section two-port driven by a source of emf, $v_{\text{in}}(t)$. The output signal is the voltage across the capacitor.*

Since the current in the circuit is

$$i(t)=C\,dv_{\text{out}}/dt$$

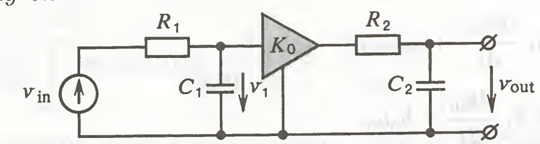
then, using the Kirchhoff voltage law, we obtain the differential equation

$$RC\,dv_{\text{out}}/dt+v_{\text{out}}=v_{\text{in}}(t) \qquad (8.32)$$

Thus, an RC-network is an example of a first-order dynamic system. The most important parameter of the circuit is the *time constant*, $\tau=RC$, which defines the time scale for the process occurring in the system.

Example 8.8. *Let there be a more complex system formed by two RC-networks decoupled by an ideal amplifier of gain K_0. It is presumed that the input impedance of the amplifier is infinitely high,*

and its output impedance is infinitesimal so the amplifier is an ideal decoupling element between circuits.



On introducing two time constants, $\tau_1 = R_1C_1$ and $\tau_2 = R_2C_2$, we obtain by analogy with the previous example the following 1st-order differential equations

$$\tau_2\, dv_{out}/dt + v_{out} = K_0 v_1$$

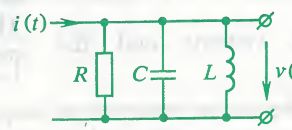
$$\tau_1\, dv_1/dt + v_1 = v_{in}(t)$$

Eliminating the dummy variable v_1 between them, we obtain the differential equation

$$\tau_1\tau_2 d^2v_{out}/dt^2 + (\tau_1 + \tau_2)\, dv_{out}/dt + v_{out} = K_0 v_{in}(t) \quad (8.33)$$

Now, we have a more complex RC-network which is a 2nd-order system.

Example 8.9. *Develop the differential equation that describes the behaviour of a lossy parallel resonant circuit:*



Here the excitation is a current $i(t)$, and the response is the voltage $v(t)$ across the resonant circuit.

On combining the currents

$$i_C = C\, dv/dt, \qquad i_L = \frac{1}{L}\int_{-\infty}^{t} v\, d\xi, \qquad i_R = v/R$$

we obtain the equation

$$C\, dv/dt + \frac{1}{L}\int_{-\infty}^{t} v\, d\xi + v/R = i(t)$$

Hence,

$$d^2v/dt^2 + 2\alpha\, dv/dt + \omega_0^2 v = (1/C)\, di/dt \quad (8.34)$$

where $\omega_0 = 1/\sqrt{LC}$ is the natural frequency of a loss-free resonant circuit, and $\alpha = 1/2RC$ is the attenuation constant of the resonant circuit.

▲ Solve Problem 4

The free (transient) response of dynamic systems. A complete analysis of the behaviour of a dynamic system described by Eq. (8.30) or Eq. (8.31) requires that we should take into account the initial conditions characterizing the internal state of the system at some fixed instant of time. It is usual to specify the values of the sought response function and of its $n-1$ derivatives for $t=0$: $u_{out}(0), \ldots, u_{out}^{(n-1)}(0)$.

From the theory of differential equations [8], it is known that the total solution of Eq. (8.31), satisfying any initial conditions, is the sum of a *particular integral* which is a particular solution of a nonhomogeneous equation (that is, one whose right-hand side is nonzero), and the *complementary function* which is the general solution of the homogeneous equation

$$a_n \frac{d^n u_{out}}{dt^n} + a_{n-1}\frac{d^{n-1}u_{out}}{dt^{n-1}} + \ldots + a_0 u_{out} = 0 \quad (8.35)$$

The solving of the homogeneous equation involves finding the roots of the characteristic equation of the system

$$a_n\gamma^n + a_{n-1}\gamma^{n-1} + \ldots + a_0 = 0 \quad (8.36)$$

■ **The property of the roots of the characteristic equation**

The characteristic equation has exactly n roots and, since the coefficients of the characteristic equation are real, the roots $\gamma_1, \gamma_2, \ldots, \gamma_n$ may be real or complex conjugate. If all roots are distinct, the complementary function, which represents the free (transient) response of (free oscillations in) the system, has the form

$$u_{free}(t) = C_1 \exp(\gamma_1 t) + C_2 \exp(\gamma_2 t) + \ldots + C_n \exp(\gamma_n t) \quad (8.37)$$

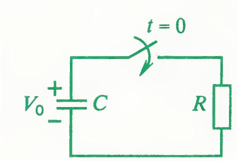
where $C_1, C_2, \ldots, C_n$ are constant numbers defined from the initial conditions. If some roots of the characteristic equation are repeated (multiple roots), the form of the complementary function is somewhat complicated owing to the appearance of the so-called *secular* factors. Then to each real root of multiplicity k we assign a set of free (transient) response functions of the form

This term has come over from astronomy

$$\exp(\gamma_i t), \quad t\exp(\gamma_i t), \ldots, t^{k-1}\exp(\gamma_i t)$$

Consider a few examples of the free (transient) response of (free oscillations in) linear stationary circuits.

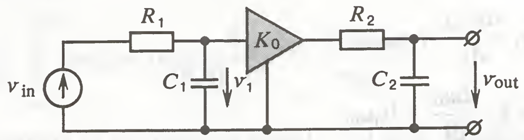
Example 8.10. *An aperiodic discharge of a capacitor. A capacitor of capacitance C, charged to a voltage V_0, is allowed to discharge into a resistor R by closing the circuit at time $t=0$. Find the manner in which the voltage across the capacitor varies.*



The system can be described by the differential equation

$$\tau\, dv_C/dt + v_C = 0$$

and its output impedance is infinitesimal so the amplifier is an ideal decoupling element between circuits.



On introducing two time constants, $\tau_1 = R_1C_1$ and $\tau_2 = R_2C_2$, we obtain by analogy with the previous example the following 1st-order differential equations

$$\tau_2\, dv_{out}/dt + v_{out} = K_0 v_1$$

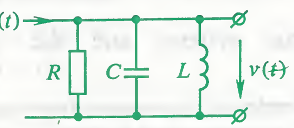
$$\tau_1\, dv_1/dt + v_1 = v_{in}(t)$$

Eliminating the dummy variable v_1 between them, we obtain the differential equation

$$\tau_1\tau_2 d^2v_{out}/dt^2 + (\tau_1+\tau_2)\, dv_{out}/dt + v_{out} = K_0 v_{in}(t) \tag{8.33}$$

Now, we have a more complex RC-network which is a 2nd-order system.

Example 8.9. *Develop the differential equation that describes the behaviour of a lossy parallel resonant circuit:*



Here the excitation is a current $i(t)$, and the response is the voltage $v(t)$ across the resonant circuit.

On combining the currents

$$i_C = C\, dv/dt, \qquad i_L = \frac{1}{L}\int_{-\infty}^{t} v\, d\xi, \qquad i_R = v/R$$

we obtain the equation

$$C\, dv/dt + \frac{1}{L}\int_{-\infty}^{t} v\, d\xi + v/R = i(t)$$

Hence,

$$d^2v/dt^2 + 2\alpha\, dv/dt + \omega_0^2 v = (1/C)\, di/dt \tag{8.34}$$

where $\omega_0 = 1/\sqrt{LC}$ is the natural frequency of a loss-free resonant circuit, and $\alpha = 1/2RC$ is the attenuation constant of the resonant circuit.

▲ **Solve Problem 4**

The free (transient) response of dynamic systems. A complete analysis of the behaviour of a dynamic system described by Eq. (8.30) or Eq. (8.31) requires that we should take into account the initial conditions characterizing the internal state of the system at some fixed instant of time. It is usual to specify the values of the sought response function and of its $n-1$ derivatives for $t=0$: $u_{out}(0), \ldots, u_{out}^{(n-1)}(0)$.

From the theory of differential equations [8], it is known that the total solution of Eq. (8.31), satisfying any initial conditions, is the sum of a *particular integral* which is a particular solution of a nonhomogeneous equation (that is, one whose right-hand side is nonzero), and the *complementary function* which is the general solution of the homogeneous equation

$$a_n \frac{d^n u_{out}}{dt^n} + a_{n-1}\frac{d^{n-1}u_{out}}{dt^{n-1}} + \ldots + a_0 u_{out} = 0 \tag{8.35}$$

The solving of the homogeneous equation involves finding the roots of the characteristic equation of the system

$$a_n\gamma^n + a_{n-1}\gamma^{n-1} + \ldots + a_0 = 0 \tag{8.36}$$

■ **The property of the roots of the characteristic equation**

The characteristic equation has exactly n roots and, since the coefficients of the characteristic equation are real, the roots $\gamma_1, \gamma_2, \ldots, \gamma_n$ may be real or complex conjugate. If all roots are distinct, the complementary function, which represents the free (transient) response of (free oscillations in) the system, has the form

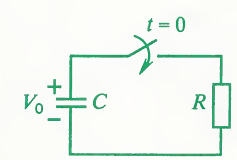
$$u_{free}(t) = C_1 \exp(\gamma_1 t) + C_2 \exp(\gamma_2 t) + \ldots + C_n \exp(\gamma_n t) \tag{8.37}$$

where $C_1, C_2, \ldots, C_n$ are constant numbers defined from the initial conditions. If some roots of the characteristic equation are repeated (multiple roots), the form of the complementary function is somewhat complicated owing to the appearance of the so-called *secular* factors. Then to each real root of multiplicity k we assign a set of free (transient) response functions of the form

This term has come over from astronomy

$$\exp(\gamma_i t), \quad t\exp(\gamma_i t), \ldots, t^{k-1}\exp(\gamma_i t)$$

Consider a few examples of the free (transient) response of (free oscillations in) linear stationary circuits.



Example 8.10. *An aperiodic discharge of a capacitor. A capacitor of capacitance C, charged to a voltage V_0, is allowed to discharge into a resistor R by closing the circuit at time $t=0$. Find the manner in which the voltage across the capacitor varies.*

The system can be described by the differential equation

$$\tau\, dv_C/dt + v_C = 0$$

subject to the initial condition

$$v_C(0) = V_0$$

The characteristic equation

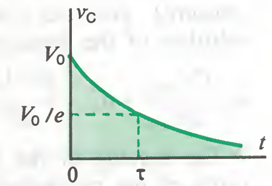
$$\tau\gamma + 1 = 0$$

has one root

$$\gamma = -1/\tau$$

The complementary function (the transient solution) can be written as

$$v_C(t) = A \exp(-t/\tau)$$



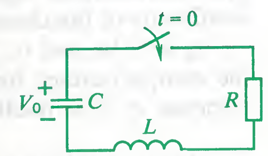
In order to satisfy the initial condition, we should set $A = V_0$. So, finally,

$$v_C(t) = V_0 \exp(-t/\tau)$$

This result defines the physical significance of the time constant τ as the time interval during which the free (transient) response decays to $1/e$ of its original value (here, $e = 2.71828\ldots$).

To sum up, the negative real root of the characteristic equation corresponds to a free (transient) response exponentially decaying with time.

Example 8.11. *An oscillatory discharge of a capacitor. The circuit of the previous example is expanded to include an inductor L.*



On the basis of the Kirchhoff second law, the differential equation of the circuit in terms of the current $i(t)$ has the form

$$d^2i/dt^2 + 2\alpha\, di/dt + \omega_0^2 i = 0 \tag{8.38}$$

where $\alpha = R/2L$ and $\omega_0 = 1/\sqrt{LC}$.

The obvious initial condition $i(0) = 0$ stems from the presence of an inductor in the circuit–the current through an inductor cannot change jumpwise. There is one more condition: At the initial instant of time the voltage across the capacitor is balanced by an emf of self-induction:

$$V_0 + L\, di/dt\,|_{t=0} = 0$$

Hence,

$$di/dt\,|_{t=0} = -V_0/L$$

The characteristic equation

$$\gamma^2 + 2\alpha\gamma + \omega_0^2 = 0$$

has complex conjugate roots

$$\gamma_{1,2} = -\alpha \pm j\sqrt{\omega_0^2 - \alpha^2} = -\alpha \pm j\omega_f$$

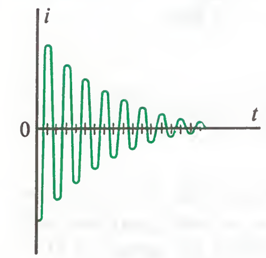
where ω_f is the frequency of the free response of (free oscillations in)

the system. As a rule, one is interested in the free response of a low-loss circuit, when $\omega_0 \gg \alpha$, so $\omega_f \approx \omega_0$.

The complementary function

$$i(t) = C_1 \exp(\gamma_1 t) + C_2 \exp(\gamma_2 t) \tag{8.39}$$

should contain the coefficients C_1 and C_2 satisfying the following set of algebraic equations (see the initial conditions):



$$C_1 + C_2 = 0$$

$$\gamma_1 C_1 + \gamma_2 C_2 = -V_0/L$$

Hence,

$$C_1 = -V_0/2j\omega_f L, \quad C_2 = V_0/2j\omega_f L$$

On substituting them in (8.39), we finally get

$$i(t) = -(V_0/\omega_f L)\exp(-\alpha t)\sin\omega_f t \tag{8.40}$$

The frequency response. If the excitation applied to a linear dynamic system is an exponential signal of the form $u_{in}(t) = \exp(j\omega t)$, then its response will be

$$u_{out}(t) = K(j\omega)\exp(j\omega t)$$

On substituting these expressions in (8.30) and cancelling out the common factor, we obtain the frequency response of the system

$$K(j\omega) = \frac{b_m(j\omega)^m + b_{m-1}(j\omega)^{m-1} + \ldots + b_1(j\omega) + b_0}{a_n(j\omega)^n + a_{n-1}(j\omega)^{n-1} + \ldots + a_1(j\omega) + a_0} \tag{8.41}$$

The frequency response of a distributed system is free from this constraint and may be described by more complex functions

Thus, the frequency response of any dynamic system described by ordinary differential equations with constant coefficients is a *rational function* of $j\omega$. The coefficients of this function are the same as the coefficients of the differential equation.

In practice, the frequency response of a linear system is usually found by inspection of the applicable circuit diagram, using the techniques of circuit theory, without writing differential equations.

Example 8.12. *For the RC-network of Example 8.7*

$$K(j\omega) = \frac{1/j\omega C}{R + 1/j\omega C} = 1/(1 + j\omega\tau) \tag{8.42}$$

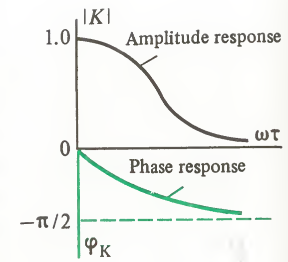
where $\tau = RC$ is the time constant.

The equation for the amplitude response takes the form

$$|K(j\omega)| = 1/\sqrt{1 + \omega^2\tau^2}$$

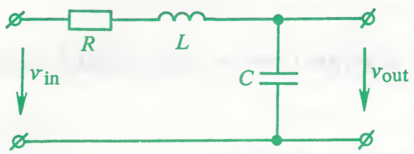
and that for the phase response

$\varphi_K(\omega) = -\arctan(\omega\tau)$



From the form of the amplitude response it is seen that the circuit in question may be used as a low-pass filter.

Example 8.13. *Analyse the frequency response of an L-section two-port consisting of an L, a C, and an R:*



Here,

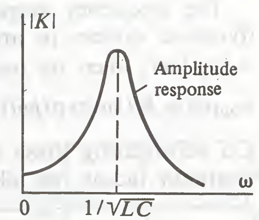
$$K(j\omega) = \frac{1/j\omega C}{R + j\omega L + 1/j\omega C} = \frac{1}{(1-\omega^2 LC) + j\omega RC}$$

Hence, the equation for the amplitude response is

$$|K(j\omega)| = \frac{1}{\sqrt{(1-\omega^2 LC)^2 + \omega^2 R^2 C^2}}$$

and that for the phase response is

$$\varphi_K(\omega) = -\arctan\frac{\omega RC}{1-\omega^2 LC}$$



If the loss resistance R is sufficiently small, so that the system Q-factor is $Q = \sqrt{L/C}/R \gg 1$, this circuit may successfully be used as a band-pass filter.

Solve Problem 5

An aperiodically loaded small-signal amplifier. An important example of linear dynamic systems is the electronic single-stage voltage amplifier shown in Fig. 8.1.

Here to make the matters more definite, the controlled electronic element is an *N-P-N* bipolar transistor. This may as well be FET (a field-effect transistor) or an electron tube.

So that any such circuits can be analysed in a unified fashion, it is customary to draw up their equivalent circuits. The equivalent-circuit method is applicable when the amplitudes of the alternating voltages around the circuit are so small that the nonlinearity of the external (load) characteristics of the devices may be neglected. For example, a bipolar transistor can be represented with sufficient accuracy by a linear equivalent circuit, if the amplitude of the a.c. component of the input voltage is small in comparison with what is known as the thermal potential of the *P-N* junction:

$v_T = kT/e$

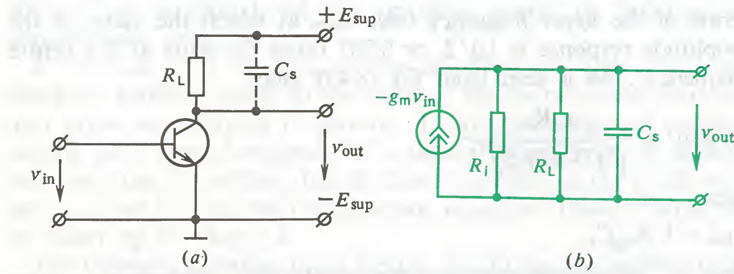


Fig. 8.1 Single-stage voltage amplifier: (*a*) simplified schematic diagram; (*b*) equivalent circuit; R_L, load resistor; C_s, stray capacitance

The stray capacitance is the sum of the output capacitance of the electron device and that of the wiring

where k is the Boltzmann constant, T is the absolute temperature of the junction, and e is the charge on an electron.

It is shown in circuit theory that the simplest equivalent circuit for the output of an active electron device (see Fig. 8.1*b*) contains a controlled current source producing a current $-g_m v_{in}$ (where g_m is the transconductance of the device at the Q-(quiescent) point), and the output (internal) resistance of the device, R_i, connected in parallel with the current source.

At $T = 300$ K (standard temperature) the thermal potential is 25 mV

The amplifier's load is a parallel combination of a load resistor R_L and a stray capacitance C_s. This type of load is known as aperiodic in contrast to an oscillatory load such as an LC-network.

The overall admittance connected in parallel with the current source is

$Y_\Sigma = 1/R_L + 1/R_i + j\omega C_s$

If the excitation applied to the amplifier is a harmonic signal at frequency ω and of complex amplitude $\dot{V}_{in}$, the complex amplitude of the output signal will be

$\dot{V}_{out} = -g_m \dot{V}_{in}/Y_\Sigma$

The "–" sign is an indication that an increase in base voltage leads to a rise in collector current and a decrease in output voltage

Hence, the frequency response of the circuit in question is

$$K(j\omega) = -g_m/Y_\Sigma = -g_m R_{eq}/(1 + j\omega C_s R_{eq}) \qquad (8.43)$$

where $R_{eq} = R_L R_i/(R_L + R_i)$.

Thus, an RC-loaded single-stage voltage amplifier has the same frequency response as the RC-network examined earlier. At zero frequency, the amplitude response function is a maximum; the magnitude of the gain factor is $K_0 = g_m R_{eq}$. As the frequency is raised, the gain falls off due to the shunting effect of the stray capacitance. The bandwidth of an amplifier is customarily stated in

The upper frequency limit (or cutoff frequency) of an amplifier

terms of the *upper frequency limit* $\omega_{\lim}$ at which the value of the amplitude response is $1/\sqrt{2}$, or 0.707 times its value at the centre frequency*. As is seen from Eq. (8.43), since

$$|K(j\omega)| = \frac{g_m R_{eq}}{\sqrt{1+\omega^2 R_{eq}^2 C_s^2}}$$

then

$$\omega_{\lim} = 1/R_{eq} C_s$$

Example 8.14. *The amplifier is set up as shown in Fig. 8.1. Its parameters are:* $R_L = 1.6\ \mathrm{k\Omega}$, $g_m = 20\ \mathrm{mAV^{-1}}$, $C_s = 30\ \mathrm{pF}$, $R_i = 15\ \mathrm{k\Omega}$. *Find the gain factor at zero frequency and the bandwidth of the system.*

To begin with, we find the equivalent load resistance:

$$R_{eq} = \frac{1.6 \times 15}{1.6 + 15} = 1.45\ \mathrm{k\Omega}$$

The magnitude of the gain factor at zero frequency is

$$K_0 = 20 \times 10^{-3} \times 1.45 \times 10^3 = 29$$

The bandwidth of the amplifier is

$$\omega_{\lim} = 1/(1.45 \times 10^3 \times 3 \times 10^{-11}) = 2.3 \times 10^7\ \mathrm{s^{-1}}$$

$$f_{\lim} = 3.66\ \mathrm{MHz}$$

Solve Problem 10

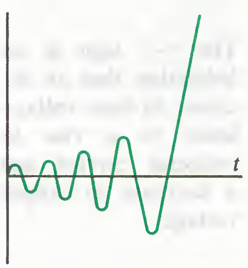
Absolute stability of a system

The stability of dynamic systems. By definition, a linear system is *absolutely stable*, if all of its natural oscillations are transients decaying with time. A necessary and sufficient condition for a dynamic system to be absolutely stable is that the real parts of all the roots of its characteristic equation, (8.36), are negative. A further condition is that the roots should not be purely imaginary. Although the free response of (transients in) the system would then be the sum of harmonic functions of the form

$$u_{\text{free}}(t) = \begin{matrix}\sin\\ \cos\end{matrix}(\omega_0 t)$$

even minute chance changes in the system parameters might drive it into an unstable state when

$$u_{\text{free}}(t) = \exp(\alpha t) \begin{matrix}\sin\\ \cos\end{matrix}(\omega_0 t),\ \alpha > 0$$



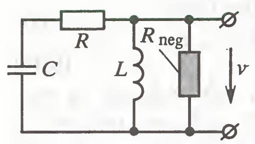
Oscillations in an unstable system

* Otherwise, it is called the *half-power bandwidth* or the *bandwidth between half-power points.– Translator's note.*

that is, the amplitude of the response would be building up exponentially with time.

If the order of a dynamic system is sufficiently high, the direct check for stability based on the roots of the characteristic equation may prove very difficult to perform. To avoid this, special stability criteria have been developed. With them, the existence of positive real roots can be verified directly from the form of the coefficients with no need to solve the characteristic equation. These criteria will be taken up in Chap. 14.

The transient response of an electric circuit can be exponentially rising only when in addition to the passive elements L, C and R it contains active elements which transfer some of the energy from external sources into the circuit. An example of active elements, known from circuit theory, is a negative-resistance resistor.



Example 8.15. *A resonant circuit whose parameters are* $C = 80\ \mathrm{pF}$, $L = 2.5\ \mathrm{\mu H}$ *and* $R = 12\ \Omega$ *contains a negative resistance connected in parallel with the inductive element. Find the critical value of the negative resistance at which the system turns unstable.*

As can be readily verified, the differential equation of the above circuit, written for the voltage v across the inductive element has the form

$$(1 + R/R_{neg})\,\mathrm{d}^2v/\mathrm{d}t^2 + (R/L + 1/R_{neg}C)\,\mathrm{d}v/\mathrm{d}t + v/LC = 0 \qquad (8.44)$$

The roots γ_1 and γ_2 of the characteristic equation have real parts

$$\mathrm{Re}\,\gamma_{1,2} = -\frac{R/L + 1/R_{neg}C}{2(1 + R/R_{neg})}$$

The system will go into an unstable state if $\mathrm{Re}\,\gamma_{1,2}$ becomes zero. Hence, the critical value of the negative resistance is given by

$$R_{neg,cr} = -L/RC = -2.604\ \mathrm{k\Omega}$$

Thus, the system in question will jump into oscillations spontaneously, if the negative resistance it contains is $R_{neg} > -2.604\ \mathrm{k\Omega}$.

8.4 Spectral (Frequency-Domain) Analysis

When speaking about spectral (frequency-domain) analysis as applied to the response of linear stationary systems, what is usually meant is a range of mathematical techniques based on the use of the properties of the frequency response. In the pages that follow, several examples will be examined in order to demonstrate the application of the spectral approach to directly finding and numerically evaluating the output signal of a system.

The basic equation. Let the excitation applied to the input of a linear stationary system be a deterministic signal $u_{\text{in}}(t)$ specified by its expansion into the Fourier integral:

$$u_{\text{in}}(t)=\frac{1}{2\pi}\int_{-\infty}^{\infty} S_{\text{in}}(\omega)\exp(j\omega t)\,d\omega \tag{8.45}$$

Suppose that the mathematical model of the system is specified by its frequency response $K(j\omega)$. As will be recalled, the complex signal $\exp(j\omega t)$ is the eigenfunction of the system, giving rise to an elementary response $K(j\omega)\exp(j\omega t)$ at the system's output. By combining the two signals, we obtain the spectral representation of the response:

$$u_{\text{out}}(t)=\frac{1}{2\pi}\int_{-\infty}^{\infty} K(j\omega)\,S_{\text{in}}(\omega)\exp(j\omega t)\,d\omega \tag{8.46}$$

This is the basic equation of the spectral method. As is seen, *the frequency response is a coefficient of proportionality between the spectra of the input and output signals*

The principle of spectral (frequency-domain) analysis

$$S_{\text{out}}(\omega)=K(j\omega)\,S_{\text{in}}(\omega) \tag{8.47}$$

Thus, a remarkable feature about system analysis in the frequency domain is that the effect of signal transformation in a system is represented simply by the algebraic operation of multiplication.

It should be stressed that the frequency-domain (spectral) approach and the time-domain approach are each other's equivalents. Indeed, the Duhamel superposition integral (8.8) is the convolution of the function $u_{\text{in}}(t)$ and the impulse response in the time domain:

$$u_{\text{out}}(t)=u_{\text{in}}(t)*h(t)$$

In consequence, the spectrum of the output signal is the product of the spectra of the functions $u_{\text{in}}(t)$ and $h(t)$. Equation (8.47) follows directly from this statement.

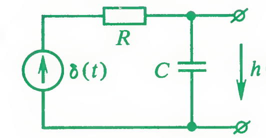
The practical value of finding the response of a system by the spectral method depends in each particular case on whether the integral in (8.46) can be evaluated.

Computation of the impulse response. As a rule, no fundamental difficulties arise in finding the frequency response of a linear system. Therefore, if it is required to find the impulse response $h(t)$ of

a system, it is advantageous to use the spectral method by which

$$h(t)=\frac{1}{2\pi}\int_{-\infty}^{\infty} K(j\omega)\exp(j\omega t)\,d\omega$$

As an example, let us find the impulse response of an RC-network whose output signal is the voltage across the capacitor. Here

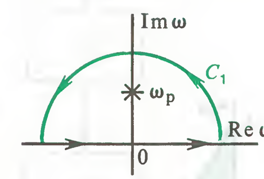


$$K(j\omega)=\frac{1}{1+j\omega RC}$$

and so the impulse response is

$$h(t)=\frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{\exp(j\omega t)\,d\omega}{1+j\omega RC} \tag{8.48}$$

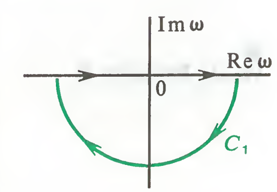
Let us employ the residue method [11] and assume that ω is a complex variable. The integration contour in (8.48) is formed by the entire real axis $\operatorname{Im}\omega=0$ and an arc C_1 of an infinitely large radius, which may close in both the upper and the lower half-plane. The integrand in (8.48) has a single simple pole at a point $\omega_p=j/RC$. The residue of the integrand at that point is



$$\operatorname{res}\left[\frac{\exp(j\omega t)}{1+j\omega RC}\right]\Bigg|_{\omega=\omega_p}=(1/jRC)\exp(-t/RC) \tag{8.49}$$

Let us find $h(t)$ for $t>0$. To this end, the arc C_1 must be located in the upper half-plane, because it is only then that the function $\exp(j\omega t)$ will tend exponentially to zero with increasing radius of the arc. In the limit, the contour integral will be equal to the integral taken solely along the real axis in accord with Eq. (8.48). By Cauchy's residue theorem, the contour integral of an analytic function is equal to $2\pi j$ multiplied by the sum of the residues of the integrand at all poles lying inside the contour of integration. Thus,

$$h(t)|_{t>0}=(1/RC)\exp(-t/RC) \tag{8.50}$$



If we want to find the impulse response for $t<0$, the contour of integration must be closed in the lower half-plane where the integrand has no poles at all, and so

$$h(t)|_{t<0}=0 \tag{8.51}$$

Graphically, the impulse response of an RC-network constructed on the basis of Eqs (8.50) and (8.51) is a curve discontinuous at $t=0$ (Fig. 8.2).

The representation of discontinuous functions by contour

▲ **Work Problems 6 and 7**

integrals is a mathematical device widely used in a variety of theoretical studies.

Computation of the output signal. As an example of using the spectral method, let us find the response of the *RC*-network defined earlier to the exponential video pulse

$$u_{in}(t) = U_0 \exp(-\alpha t)\,\sigma(t)$$

Now the spectrum of the excitation is

$$S_{in}(\omega) = U_0/(\alpha + j\omega)$$

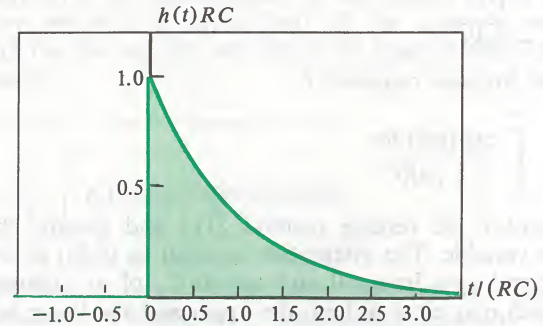
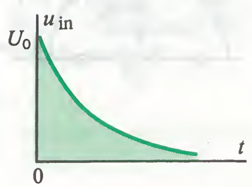


Fig. 8.2 Plot of the impulse response of an *RC*-network



and the problem reduces to evaluating the integral

$$u_{out}(t) = (U_0/2\pi\alpha)\int_{-\infty}^{\infty}\frac{\exp(j\omega t)\,d\omega}{(1+j\omega/\alpha)(1+j\omega RC)} \tag{8.52}$$

On expanding the integrand into partial fractions, we get

$$\frac{1}{(1+j\omega/\alpha)(1+j\omega RC)} = \frac{1}{1-\alpha RC}\left(\frac{1}{1+j\omega/\alpha} - \frac{\alpha RC}{1+j\omega RC}\right)$$

Since the terms in the parentheses have the same structure, we may directly use the results obtained in finding the impulse response and write the solution as

▲ **Work Problem 8**

$$|u_{out}(t)|_{t>0} = \frac{U_0}{1-\alpha RC}[\exp(-\alpha t) - \exp(-t/RC)] \tag{8.53}$$

Naturally,

$$u_{out}(t)|_{t<0} = 0 \tag{8.54}$$

The applicable plot appears in Fig. 8.3.

The frequency response of a multistage system. Characteristically, telecommunications use complex systems in which the individual sections are connected in cascade so that the

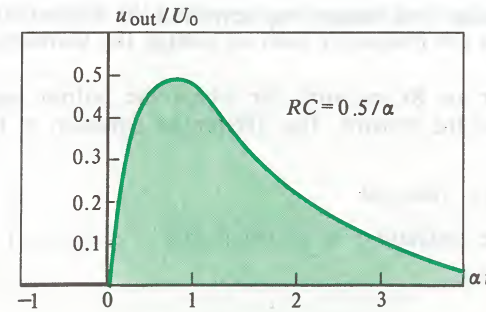
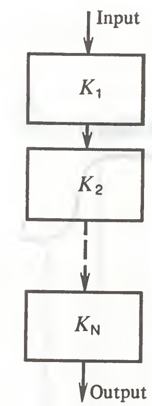


Fig. 8.3 Response of an *RC*-network to an exponential video pulse. (Note that the network smoothes the input signal.)

output from a previous stage acts as the input signal for the next. An example is a multistage amplifier.



Suppose that we know the frequency responses of the individual stages $K_n(j\omega)$ ($n = 1, 2, \ldots, N$), where N is the total number of stages. If the excitation for the first stage is

$$u_{in}(t) = \exp(j\omega t)$$

its output signal will be

$$u_{out}(t) = K_1(j\omega)K_2(j\omega)\ldots K_N(j\omega)\exp(j\omega t)$$

Hence, the overall frequency response of the system is

$$K_{overall}(j\omega) = \prod_{n=1}^{N} K_n(j\omega) \tag{8.55}$$

● **Decibels**

In engineering practice, the amplitude (magnitude) response of systems is ordinarily expressed in *decibels*, special logarithmic units. If at any frequency ω we know the magnitude of the frequency response, then the *gain* of the system at that frequency, as expressed in decibels, will be

$$\Delta = 20\log_{10}|K(j\omega)| \tag{8.56}$$

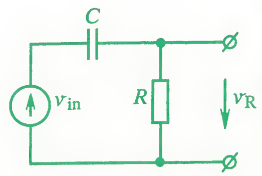
A system for which $|K(j\omega)| < 1$ will attenuate the signal, and instead of a gain we have a loss, or negative gain.

It is an easy matter to see that when several stages are connected in cascade, their gains are added together:

$$\Delta_{overall} = \sum_{n=1}^{N}\Delta_n \tag{8.57}$$

Differentiating and integrating networks. In telecommunications, linear circuits are frequently used to change the waveform of pulse signals.

Consider an RC-network for which the output signal is the voltage across the resistor. The differential equation of this system is



$$\tau\, dv_R/dt + v_R = \tau\, dv_{in}/dt \tag{8.58}$$

If the time constant τ is so small that

$$\tau\,|dv_R/dt| \ll |v_R| \tag{8.59}$$

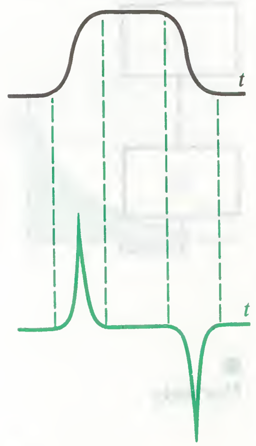
at any instant of time, we may neglect the first term on the left-hand side of Eq. (8.58) in comparison with the second, and so

$$v_R \approx \tau\, dv_{in}/dt \tag{8.60}$$

that is, if the condition (8.59) is satisfied, the RC-network performs the approximate differentiation of the applied signal. In practical applications, differentiating networks are used for pulse peaking (or sharpening).

Clearly, whether the inequality (8.59) is satisfied or not depends not only on the circuit parameters, but also on the characteristics of the input signal. For estimation purposes, it is simpler and more instructive to use analysis in the frequency domain. The frequency response of the circuit in question

The input signal



and the output signal of a differentiating network

$$K(j\omega) = \frac{j\omega\tau}{1 + j\omega\tau}$$

will be close to that of an ideal differentiator

$$K(j\omega) \approx j\omega\tau$$

if the product $\omega\tau$ is negligible in comparison with unity in that frequency region where the bulk of the signal energy is concentrated. As an example, let the input signal be a rectangular video pulse of duration τ_p. Using a rough estimate for the upper frequency limit (the bandwidth) of the pulse

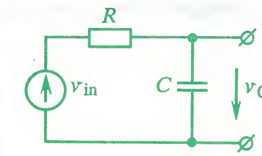
$$\omega_{lim} = 2\pi/\tau_p$$

we obtain the condition for the applicability of an RC-network for the approximate differentiation of such a signal

$$\tau = RC \ll \tau_p/2\pi \tag{8.61}$$

If in an RC-network the output signal is picked off the capacitor, the situation will be the reverse of the previous case. Here,

$$\tau\, dv_C/dt + v_C = v_{in}(t)$$



and, if the circuit and excitation parameters are such that

$$\tau\,|dv_C/dt| \gg |v_C|$$

then

$$v_C(t) \approx \frac{1}{\tau}\int_{-\infty}^{t} u_{in}(\xi)\, d\xi \tag{8.62}$$

An RC-network possessing the above properties is called an integrating network.

The accuracy of approximate integration improves as the fraction of h.f. components in the spectrum of the excitation increases. To demonstrate, since here

$$K(j\omega) = 1/(1 + j\omega\tau)$$

then the approximate equality

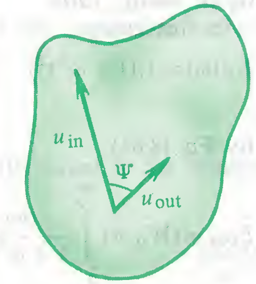
$$K(j\omega) \approx 1/j\omega\tau$$

▲ **Solve Problem 9**

assuring the integrating properties of the network will take place if $\omega_l\tau \gg 1$, where ω_l is the lower frequency limit of the signal bandwidth.

Integrating networks capable of suppressing the h.f. components of the input signal are often used as smoothing filters. Also, they can convert stepwise changes in the input signal (step inputs) into a linearly rising (or ramp) output voltage.

Geometric interpretation of signal transformation in a linear system. The spectral method provides a tool for an easy-to-grasp analysis of the transformations that a signal experiences in passing through a linear stationary system. From the view-point of the geometric concepts developed in Chap. 1, the system operator T defines the law by which one signal $u_{in}(t)$ in some linear space is transformed, or mapped, into a new signal $u_{out}(t)$. As a rule, this functional space is a Hilbert space. Then, in the most general case it may be argued that the operator T changes the norm of the signal $u_{in}(t)$, that is,



$$\|u_{in}\| \neq \|Tu_{in}\|$$

Also, an angle ψ is formed between the signals u_{in} and u_{out}.

By the Rayleigh formula (see Chap. 3), the energy of the output

signal is given by

$$E_{out} = \| u_{out} \|^2 = \frac{1}{2\pi}\int_{-\infty}^{\infty} S_{out}(\omega) S^*_{out}(\omega)\, d\omega$$

$$= \frac{1}{\pi}\int_{0}^{\infty} | K(j\omega)|^2 W_{in}(\omega)\, d\omega \qquad (8.63)$$

where $W_{in}(\omega)$ is the power spectrum of the input signal.

In accordance with (8.63),

$$W_{out}(\omega) = | K(j\omega)|^2 W_{in}(\omega)$$

The quantity

The power transfer response

$$K_P(\omega) = | K(j\omega)|^2 \qquad (8.64)$$

is called the *power transfer response* at the specified frequency ω. Since this function is real, the energy of the output signal is far simpler to find than the exact form of the output signal. It is to be noted that in many cases it is quite enough to know only the manner in which the energy of the signal passing through a linear system varies.

Example 8.16. *Let an RC-network with a frequency response $K(j\omega) = 1/(1 + j\omega\tau)$ be driven by an ideal low-pass signal whose power spectrum is W_0 in the frequency interval $0 < \omega < \omega_c$ (where ω_c is the cut-off frequency) and zero elsewhere. Find the input to output signal energy ratio.*

In this case,

$$K_P(\omega) = 1/(1 + \omega^2\tau^2)$$

$$K_P(\omega) = K(j\omega)\, K(-j\omega)$$

By Eq. (8.63)

$$E_{out} = (W_0/\pi)\int_0^{\omega_c} \frac{d\omega}{1 + \omega^2\tau^2} = (W_0/\pi\tau)\arctan \omega_c\tau$$

The energy of the input signal is

$$E_{in} = W_0\omega_c/\pi$$

Therefore, the energy ratio

$$\eta = E_{out}/E_{in} = \arctan \omega_c\tau/(\omega_c\tau) \qquad (8.65)$$

tends to zero as both the time constant τ and the cut-off frequency ω_c are increased.

The angle between the input and output signal phasors. In Chap. 1 it has been shown how two signals can be compared by finding the angle ψ between the signal phasors in the Hilbert space. This idea can be utilized for comparing the input and output signals of a linear stationary system.

By the generalized Rayleigh theorem, the scalar product of the two signals can be expressed in terms of their spectra

Solve Problem 14

$$(u_{in},\ u_{out}) = \frac{1}{2\pi}\int_{-\infty}^{\infty} S_{in}(\omega) S^*_{out}(\omega)\, d\omega$$

$$= \frac{1}{2\pi}\int_{-\infty}^{\infty} S_{in}(\omega) S^*{}_{in}(\omega) K^*(j\omega)\, d\omega$$

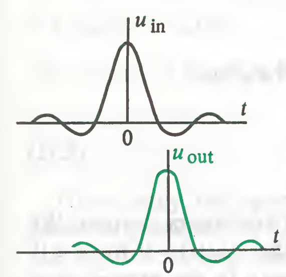
$$= \frac{1}{2\pi}\int_{-\infty}^{\infty} W_{in}(\omega) K^*(j\omega)\, d\omega$$

Since the imaginary part of the frequency response is an odd function of frequency, the above equation can be simplified as follows:

$$(u_{in},\ u_{out}) = \frac{1}{\pi}\int_{0}^{\infty} W_{in}(\omega)\, \mathrm{Re}\, K(j\omega)\, d\omega \qquad (8.66)$$

The angle ψ between the phasors of the input and output signals can be found from the relation

$$\cos\psi = \frac{(u_{in},\ u_{out})}{\| u_{in} \| \cdot \| u_{out} \|} \qquad (8.67)$$



Example 8.17. *Find the angle ψ between the input and output signals in the RC-network of Example 8.16.*

Since here

$$\mathrm{Re}\, K(j\omega) = 1/(1 + \omega^2\tau^2)$$

it follows that the integral (8.66) is numerically equal to the square of the norm of the output signal. Hence,

$$\cos\psi = \sqrt{\eta} = \left(\frac{\arctan \omega_c\tau}{\omega_c\tau}\right)^{1/2} \qquad (8.68)$$

If the product $\omega_c\tau \gg 1$, then $\cos\psi \to 0$, and this implies that the output signal of the *RC*-network is nearly orthogonal with respect to the input signal. An insight into this effect can be obtained from quantitative considerations, if we recall that the network delays the output signal owing to the lag in its response.

The autocorrelation characteristic of a system. Before we conclude our overview of spectral (frequency-domain) analysis as it is applied in the theory of stationary systems, one more useful function must be mentioned. This is the autocorrelation characteristic of a system, $\Xi(\tau)$. Customarily, it is defined as the Fourier transform of the power transfer function

$$\Xi(\tau)=\frac{1}{2\pi}\int_{-\infty}^{\infty} K_P(\omega)\exp(j\omega t)\,d\omega \tag{8.69}$$

Alternatively, the function in (8.69) may be expressed in the time domain. To this end, we note that

$$K_P(\omega)=K(j\omega)K^*(j\omega)$$

Therefore the relation between $K_P(\omega)$ and $\Xi(\tau)$ must be the same as has been found (in Chap. 3) to exist between the power spectrum and the autocorrelation function of an arbitrary signal

$$\Xi(\tau)=\int_{-\infty}^{\infty} h(t)h(t-\tau)\,dt \tag{8.70}$$

8.5 The Operational Method

There is a widely used operational method of analysis closely related to the spectral method just discussed. It is based on representing input and output signals by their Laplace transforms.

In this book the term "operational method" refers solely to the Laplace transformation

Solving differential equations by the operational method. This method is an exceptionally flexible and powerful tool enabling one to solve linear differential equations with constant coefficients by formalized procedures. Precisely this property has made it so popular in the study of linear dynamic systems.

Let the differential equation

$$a_n\frac{d^n u_{out}}{dt^n}+a_{n-1}\frac{d^{n-1}u_{out}}{dt^{n-1}}+\ldots+a_1\frac{du_{out}}{dt}+a_0u_{out}$$
$$=b_m\frac{d^m u_{in}}{dt^m}+b_{m-1}\frac{d^{m-1}u_{in}}{dt^{m-1}}+\ldots+b_0u_{in} \tag{8.71}$$

define the input-output relation for a linear stationary system. We impose an important constraint: We deem that $u_{in}(t)=0$ for $t<0$. Also, in view of the class of problems dealt with in the present text, we deem that the system does not contain any stored energy up to the instant when the input signal occurs. Mathematically this means that we should choose zero initial conditions:

$$u_{out}(0)=u'_{out}(0)=\ldots=u_{out}^{(n-1)}(0)=0$$

Finally, we assume that the area of feasible input signals does not contain any functions so rapidly increasing with time that no Laplace transform exists for them.

The correspondence between the original signal and its Laplace transform will be designated as follows:

$$u_{in}(t)\risingdotseq U_{in}(p),\quad u_{out}(t)\risingdotseq U_{out}(p)$$

On taking the Laplace transforms of both sides of Eq. (8.71), we get

$$(a_np^n+a_{n-1}p^{n-1}+\ldots+a_1p+a_0)U_{out}(p)$$
$$=(b_mp^m+b_{m-1}p^{m-1}+\ldots+b_1p+b_0)U_{in}(p) \tag{8.72}$$

The most important characteristic that forms the basis of the Laplace transform method is the ratio of the transform of the output signal to the transform of the input signal:

$$K(p)=U_{out}(p)/U_{in}(p) \tag{8.73}$$

called the *transfer function* of a system.

The transfer function of a system

In accordance with Eq. (8.72),

$$K(p)=\frac{b_mp^m+b_{m-1}p^{m-1}+\ldots+b_1p+b_0}{a_np^n+a_{n-1}p^{n-1}+\ldots+a_1p+a_0} \tag{8.74}$$

Within the framework of the Laplace transformation, the transfer function is a complete mathematical model of the system. If the transfer function of a system is known, the response of the system to a specified excitation can be found by a procedure which can be broken down into three steps:

(1) $u_{in}(t)\to U_{in}(p)$

(2) $U_{out}(p)=K(p)U_{in}(p)$

(3) $U_{out}(p)\to u_{out}(t)$

Historically, the operational method goes back to the operational calculus proposed by Heaviside as far back as the end of the 19th century for solving differential equations describing non-stationary processes in linear electric circuits. The Heaviside method is based on the formal replacement of the differentiation operator d/dt with a complex number p. (For connections between the Heaviside and Laplace transform methods the reader is referred to [11].)

Properties of the transfer function. Comparison of Eqs. (8.74) and (8.41) will show that the function $K(p)$ is an analytic continuation of the frequency response $K(j\omega)$ from the imaginaly $(j\omega)$ axis to the

entire plane of the complex frequency $p = \sigma + j\omega$. The function $K(p)$ is analytic over the entire p-plane, except for a finite number of points $p_1, p_2, \ldots, p_n$, which are the roots of the denominator in (8.74). These points, that is, the roots of the equation

$$a_n p^n + a_{n-1} p^{n-1} + \ldots + a_1 p + a_0 = 0$$

are called the *poles* of the transfer function $K(p)$.

The set of points $z_1, z_2, \ldots, z_m$, which are the roots of the equation

$$b_m z^m + b_{m-1} z^{m-1} + \ldots + b_1 z + b_0 = 0$$

are the *zeros* of the transfer function.

Poles and zeros

On placing out the common multiplier K_0 which arises when the numerator is divided into the denominator in Eq. (8.74), we may write $K(p)$ in the so-called pole-zero representation:

$$K(p) = K_0 \frac{(p - z_1)(p - z_2)\ldots(p - z_m)}{(p - p_1)(p - p_2)\ldots(p - p_n)}$$

Since the coefficients of the differential equation (8.72) are real, the poles and zeros have an important property: *They are all either real or form complex conjugate pairs.*

The usual practice is to show poles as stars and zeros as circles

Frequently, the transfer function is depicted graphically by means of the so-called pole-zero diagram which is a plane where the coordinates of the points are marked by suitable symbols. The transfer function $K(p)$, being complex, cannot in itself be presented in graphic form. Therefore, the usual procedure is to show the three-dimensional surface of the function $|K(p)|$ over a plane with Cartesian coordinates (Fig. 8.4).

The surface has a typical "mountainous" appearance, with the infinite peaks corresponding to the poles of the transfer function, and with the valleys to its zeros. On cutting the surface with a plane which contains both the vertical axis and the $j\omega$-axis, we obtain the amplitude response profile of the system.

Work Problem 11

The poles of the transfer function for a linear circuit are nothing but the roots of the characteristic equation (8.36). Therefore, for a system to be absolutely stable, it is necessary and sufficient that these poles lie only in the left-hand half of the p-plane. In the general case, the zeros of the transfer function may be located in both the left-hand and the right-hand half of the p-plane.

The Laplace inversion formula. The final step in finding the response of a linear system to a signal by the Laplace transform method is the recovery of the original time function for which the

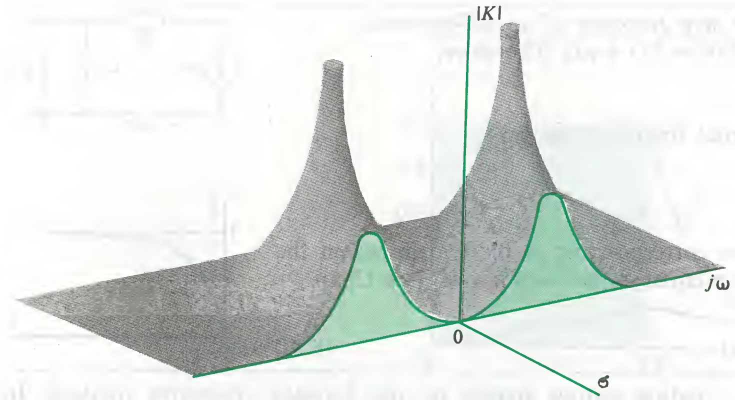


Fig. 8.4 The appearance of the $K(p)$ surface for a transfer function having two complex conjugate poles: $p_{1,2} = -\alpha \pm j\omega_0$, and one zero: $z = 0$

Laplace transform is

$$U_{out}(p) = K(p)\, U_{in}(p)$$

Consider the special case when the function $U_{out}(p)$ is the ratio of two polynomials in powers of the complex frequency:

$$U_{out}(p) = M(p)/N(p)$$

such that the power m of the numerator does not exceed the power n of the denominator and also the roots of the denominator p_i ($i = 1, 2, \ldots, n$) are simple.

The method by which the original time function can be recovered from its Laplace transform is based on the representation of $U_{out}(p)$ as a sum of partial fractions:

$$U_{out}(p) = \sum_{i=1}^{n} C_i \frac{1}{p - p_i}$$

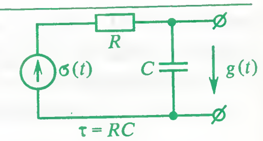
The coefficients C_i are the residues of the function $U_{out}(p)$ at poles, therefore [11]

$$U_{out}(p) = \sum_{i=1}^{n} \frac{M(p_i)}{N'(p_i)} \frac{1}{p - p_i} \tag{8.75}$$

As will be recalled, $1/(p - p_i)$ is the Laplace transform of $\exp(p_i t)$. Hence, the following Laplace inversion formula results from (8.75):

$$u_{out}(t) = \sum_{i=1}^{n} \frac{M(p_i)}{N'(p_i)} \exp(p_i t) \tag{8.76}$$

Example 8.18. *Find the step response of an RC-network.*

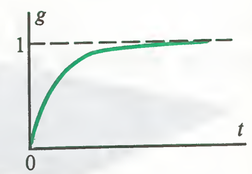


Here, $\sigma(t) \doteq 1/p$ and $K(p) = 1/(1 + p\tau)$. Therefore,

$$U_{\text{out}}(p) = 1/p(1 + p\tau)$$

On expanding into partial fractions, we get

$$U_{\text{out}}(p) = \frac{1}{p} - \frac{1}{p + 1/\tau}$$

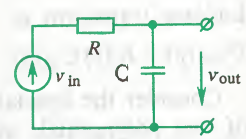


The original time functions corresponding to the two terms on the right-hand side of the above equation are well known (see Chap. 2). The result is

$$g(t) = [1 - \exp(-t/\tau)]\,\sigma(t) \tag{8.77}$$

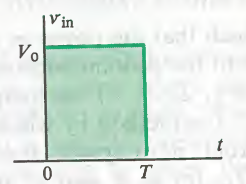
Finding output signals by the Laplace transform method. In using the Laplace transform method, the bulk of the computational work can be avoided by reference to the widely known tables of Laplace transform pairs (see Appendix 4).

Example 8.19. *The response of an RC-network to a rectangular video pulse. Let the excitation applied to the RC-network shown in the accompanying figure be a rectangular voltage video pulse of the known parameters V_0 and T. Find the function describing the output signal.*



The Laplace transform of the input signal is

$$V_{\text{in}}(p) = (V_0/p)[1 - \exp(-pT)]$$



The term $\exp(-pT)$ implies that there is a time shift by T. Therefore, on the basis of the results obtained in Example 8.18, we may write

$$v_C(t) = V_0(1 - e^{-t/\tau})\sigma(t) - V_0[1 - e^{-(t-T)/\tau}]\sigma(t - T) \tag{8.78}$$

It is more instructive to re-write (8.78) as

$$v_C(t) = V_0(1 - e^{-t/\tau}) \text{ for } 0 < t < T$$

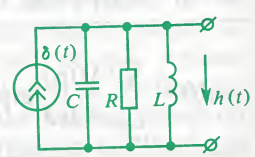
$$v_C(t) = -V_0 e^{-t/\tau}(1 - e^{T/\tau}) \text{ for } t > T \tag{8.79}$$

If the output signal is picked off the resistor, then for the same values of R and C, the voltage across the resistor will be

$$v_R(t) = v_{\text{in}} - v_C(t)$$

The respective plots appear in Figs. 8.5 and 8.6.

Example 8.20. *The impulse response of a parallel resonant circuit. Let a lossy parallel resonant circuit be driven by a delta impulse of current in the common part of the circuit. The output signal is the voltage across the circuit.*



The equality $V(p) = Z(p)I(p)$ indicates that in this case the

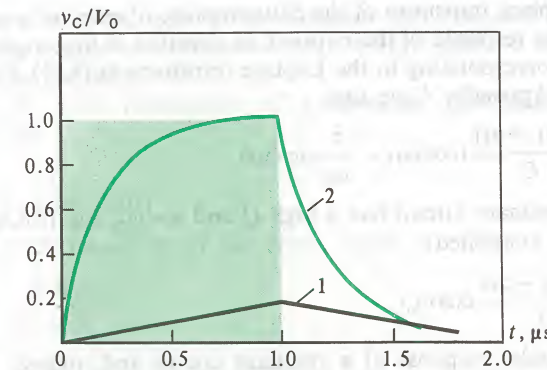


Fig. 8.5 Time variations in the voltage across the capacitor of an *RC*-network driven by a rectangular video pulse of $T = 1\mu$s duration: (*1*) for $T/\tau = 0.2$; (*2*) for $T/\tau = 5$

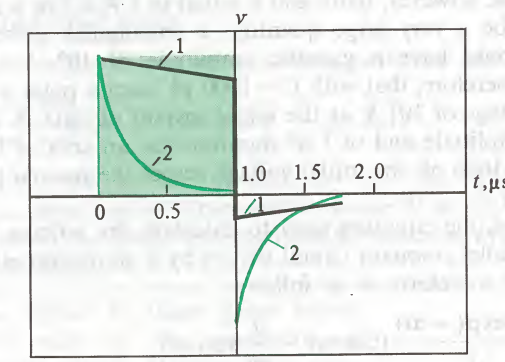


Fig. 8.6 Time variations in the voltage across the capacitor of an *RC*-network driven by a rectangular video pulse of $T = 1\mu$s duration: (*1*) for $T/\tau = 0.2$; (*2*) for $T/\tau = 5$

transfer function is the Laplace transform of the resonant-circuit impedance

$$Z(p) = \frac{pRL}{p^2LCR + pL + R} = \frac{p/C}{p^2 + 2\alpha p + \omega_0^2} \tag{8.80}$$

where $\alpha = 1/2RC$ and $\omega_0^2 = 1/LC$.

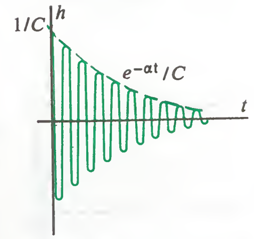
It is convenient to re-write Eq. (8.80) as

$$Z(p) = \frac{p/C}{(p + \alpha)^2 + \omega_f^2} \tag{8.81}$$

where $\omega_f = \sqrt{\omega_0^2 - \alpha^2}$ is the natural frequency of the lossy resonant circuit (the frequency of its free response).

The Laplace transform of the delta-impulse of current is unity, so the impulse response of the network in question is the original time function corresponding to the Laplace transform in (8.81). From the tables in Appendix 4 we find

$$h(t)=\frac{\exp(-\alpha t)}{C}\left(\cos\omega_f t-\frac{\alpha}{\omega_f}\sin\omega_f t\right) \tag{8.82}$$



If the resonant circuit has a high Q and $\alpha \ll \omega_0$, Eq. (8.82) can be somewhat simplified:

$$h(t)\approx\frac{\exp(-\alpha t)}{C}\cos\omega_0 t \tag{8.83}$$

The impulse response of a resonant circuit and, indeed, that of any other linear oscillatory system has the typical form of a harmonic wave with an envelope exponentially decreasing in time.

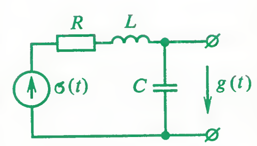
It is important to remember that Eqs. (8.82) and (8.83) hold when a resonant circuit is driven by an infinitesimally short current pulse whose area is, however, finite and is equal to 1 A s. On a real scale, this would be a very large quantity – a rectangular pulse of 1 μs duration would have a gigantic amplitude of 10^6 A. It is not surprising, therefore, that with $C = 1000$ pF such a pulse would give rise to a voltage of 10^9 V at the initial instant of time. A real pulse of 0.01 A amplitude and of 1 μs duration has an area of 10^{-8} A s, and for $C = 1000$ pF the initial voltage across the resonant circuit is a mere 10 V.

To sum up, the equation used to calculate the voltage produced across a parallel resonant circuit driven by a short current pulse of an arbitrary waveform is as follows:

$$v_{out}(t)=\frac{A_p\exp(-\alpha t)}{C}\left(\cos\omega_f t-\frac{\alpha}{\omega_f}\sin\omega_f t\right) \tag{8.84}$$

where A_p is the area of the pulse.

Example 8.21. *The step response of a series resonant circuit.* Here, the transfer function is



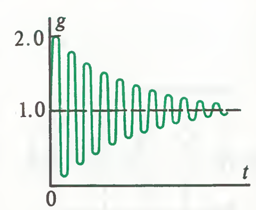
$$K(p)=\frac{1/pC}{pL+R+1/pC}=\frac{\omega_0^2}{p^2+2\alpha p+\omega_0^2}$$

where $\alpha = R/2L$. The excitation is the Heaviside function for which the Laplace transform is $1/p$. Hence,

$$V_{out}(p)=\frac{\omega_0^2}{p\left[(p+\alpha)^2+\omega_f^2\right]}$$

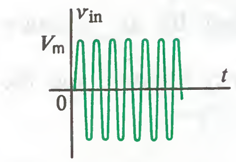
From the table of Laplace transform pairs, we finally get

$$g(t)=1-e^{-\alpha t}\left(\cos\omega_f t+\frac{\alpha}{\omega_f}\sin\omega_f t\right) \tag{8.85}$$



As is seen from the plot of the function $g(t)$, the unit step input causes the system to oscillate so that it approaches a new steady state asymptotically.

Example 8.22. *Connection of a harmonic emf source to an RC-network.*

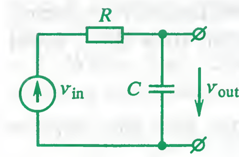


Let

$$v_{in}(t)=V_m\sin\omega_0 t\,\sigma(t)$$

The Laplace transform of the above signal is

$$V_{in}(p)=\frac{V_m\omega_0}{p^2+\omega_0^2}$$

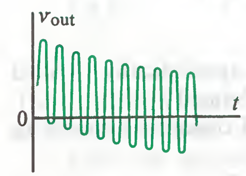


Since

$$K(p)=1/(1+p\tau)=\alpha/(p+\alpha)$$

where $\alpha = 1/\tau$, it follows that

$$V_{out}(p)=\frac{V_m\alpha\omega_0}{(p+\alpha)(p^2+\omega_0^2)}$$



From the table of Laplace transform pairs we find

$$V_{out}(t)=\frac{V_m\alpha\omega_0}{\omega_0^2+\alpha^2}\left(e^{-\alpha t}-\cos\omega_0 t+\frac{\alpha}{\omega_0}\sin\omega_0 t\right) \tag{8.86}$$

The first term in the parentheses on the right-hand side represents a decaying free (transient) response. If $\alpha t \gg 1$, the solution will only retain the forced (or steady-state) response which varies in time harmonically.

Many more examples could be added to those given above, including more complicated ones such as the connection of a harmonic emf source to a resonant circuit. However, the exact solutions that are obtained are fairly unwieldy. It appears more convenient to carry out the analysis by the method applicable to nonstationary processes in oscillatory circuits, set forth in Chap. 9.

Summary

- The relation connecting the input and output signals of a system is called the system operator.
- Systems are classed according to the properties of system operators. Systems may be linear or nonlinear, stationary or nonstationary, lumped-constant or distributed-constant.
- The response of a linear system to a delta impulse is called the impulse response of that system.
- The output signal is the convolution of the input signal and the impulse response. The frequency response function and the impulse response function are related by a Fourier transform pair.

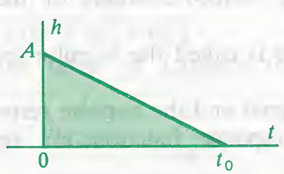
- The free (transient) response of a dynamic system is defined by the roots of the characteristic equation.
- A dynamic system is absolutely stable, if all the roots of the characteristic equation have negative real parts.
- The frequency response of a linear dynamic system described by an ordinary differential equation is a rational function of frequency.
- The spectrum of the output signal is the product of the frequency response and the spectrum of the input signal.

Review Questions

1. Give several examples of linear and nonlinear, stationary and nonstationary systems.
2. What are the conditions under which the response of a linear system to a short input pulse may be represented by the impulse response of the system?
3. State the condition for the physical realizability of a system.
4. Define the step response of a system. How are the step response and the impulse response related?
5. Define the frequency response function of a linear stationary system.
6. Formulate the Paley-Wiener criterion.
7. What is the salient property of dynamic systems?
8. Write the equation defining the frequency response of an aperiodically loaded small-signal amplifier. How is its cut-off frequency (bandwidth) defined?
9. What is the essence of the spectral (frequency-domain) analysis of dynamic systems as applied to their response?
10. What logarithmic units are used to express the gain of a system?
11. Draw up schematic circuit diagrams for a differentiating network and an integrating network and explain their principle of operation.
12. How does a linear circuit transform the input signal phasor regarded as an element of the Hilbert space?
13. What is the power transfer function?
14. Define the transfer function of a linear stationary system.

Problems

1. The impulse response $h(t)$ of a linear stationary system is a triangular pulse:



The circuit is excited by the signal

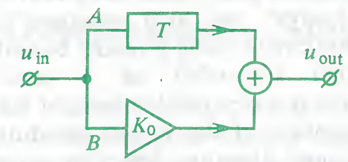
$$v_{\text{in}}(t)=\begin{cases} at, & t>0\\ 0, & t<0\end{cases}$$

Find the response of the system.

2. Calculate the impulse response of an ideal integrator for which

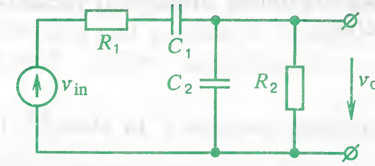
$$v_{\text{out}}(t)=\int_0^t v_{\text{in}}(\xi)\,d\xi$$

3. The block diagram of a system has the form



The arm A contains a delay unit which delays the input signal for a time T, and the arm B a scaler (amplifier) of gain K_0. Find the impulse response of the system.

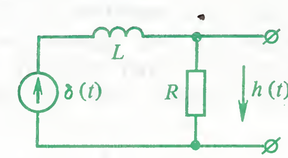
4. Write the differential equation describing the following circuit:



The equation must be written for the unknown function $v_{\text{out}}(t)$.

5. Find the frequency response functions for the systems in Problems 3 and 4.

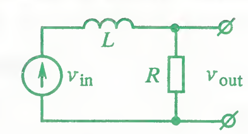
6. Calculate the impulse response of the RL-network whose circuit diagram has the form



7. Calculate the impulse response of the circuit examined in Problem 4.

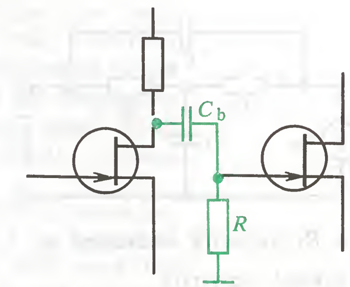
8. Analyse Eq. (8.53) for $\alpha=1/RC$.

9. Establish the conditions under which the circuit shown in the accompanying diagram:



can serve as a network taking an approximate integral of the input signal.

10. The multistage FET amplifier shown in the accompanying diagram contains a d.c. blocking capacitor C_b:



As its name implies, its function is to block the transfer of the high d.c. potential from the drain of the previous stage to the gate of the next. The amplifier is intended to amplify rectangular video pulses of duration $T=1$ ms. Considering that the resistor value is $R=0.5$ MΩ, and the FETs have an infinite input impedance, determine the value of C_b for which the gate voltage (see the schematic) drops by not more than 5% from its maximum level over the pulse duration T.

11. Find the transfer function of a two-stage amplifier consisting of identical RC-loaded stages. The parameters of one stage are $g_m=10$ mA V^{-1}, $R=0.3$ kΩ, $R_i=7$ kΩ, and $C_b=120$ pF. The capacitance of C_b (see Problem 10) is so high that its effect on the amplifier performance may be neglected.

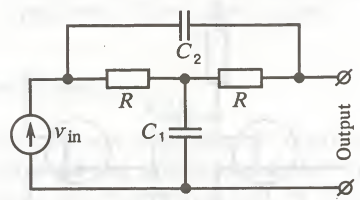
12. A series LCR resonant circuit is excited by

$$v_{\text{in}}(t)=\alpha t\sigma(t)$$

Derive the relations defining the manner in which the voltages across the capacitor and the inductor vary.

Advanced Problems

13. Find the frequency response and the impulse response of the circuit shown below:



14. An RC-network is excited by

$v_{in}(t) = V_0 \exp(-\alpha t)\,\sigma(t)$

The output signal is picked off the capacitor. Determine the angle between the input and output signal phasors in the Hilbert space.

15. Analyse the step response of an oscillator circuit using a model consisting of a weight suspended on a cord. The excitation is a step displacement of the point of suspension of the weight (pendulum) in a horizontal direction. Select by experiment the form of excitation that would move the system from one steady state to another in a finite time. Draw the conclusion about the limiting speed of response of the oscillatory system. (This problem is an illustration to the theory of optimal control [42] which is a rapidly expanding division of present-day cybernetics.)

Chapter 9

Response of Frequency-Selective Systems to Deterministic Signals

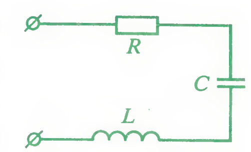
Since their early days, telecommunication systems have been widely using frequency-selective linear circuits in order to extract wanted signals. Circuits in this class attenuate all frequencies except those lying in the relatively narrow band around the centre frequency. Frequency filtering is effective when the signal being processed is sufficiently narrowband. Examples are the various modulated signals examined in Chap. 4.

Narrowband frequency-selective circuits or, as they are more frequently called, linear narrowband frequency filters, have a number of specific properties. In this chapter the reader will learn the techniques and procedures developed for the analysis and synthesis of such circuits, including their characteristics and responses.

9.1 Models of Frequency-Selective Circuits

The simplest frequency filter is a resonant circuit formed by appropriately interconnecting an inductor L, a capacitor C, and a resistor R. Series as well as parallel resonant circuits are dealt with in a course on circuit theory. Without going into a detailed derivation which is presumed to be known to the reader [25] we will re-state the basic points which will be frequently used in the subsequent discussion.

Frequency response of a series resonant circuit. If we take the input as a voltage and the output as a current, the behaviour of a series resonant circuit driven by a harmonic excitation will be described by the *complex input* (or *driving-point*) admittance



$$Y(j\omega) = \frac{1}{R + j\omega L + 1/j\omega C} \tag{9.1}$$

At the resonance frequency, $\omega_{res} = 1/\sqrt{LC}$, the admittance of the resonant circuit is purely conductive

$Y_{res} = 1/R$

The quantity

$$\rho = \sqrt{L/C} \tag{9.2}$$

is called the *characteristic impedance* of the resonant circuit.

The frequency-selective properties of this system improve with an

increase in its *Q-factor* defined as

$$Q = \rho/R = \omega_{res} L/R = 1/\omega_{res} RC \tag{9.3}$$

It is customary to introduce the *absolute detuning* $\Delta\omega$ of the input harmonic signal relative to the resonance frequency by writing the source frequency as

$$\omega = \omega_{res} + \Delta\omega$$

and also the dimensionless *relative detuning*

$$\nu = \omega/\omega_{res} - \omega_{res}/\omega \tag{9.4}$$

which vanishes at the resonant frequency.

Generalized detuning

Finally, there is the *generalized detuning* ξ which is defined as the ratio of the reactive impedance of the resonant circuit to the loss resistance:

$$\xi = \frac{1}{R}(\omega L - 1/\omega C) = Q\nu \tag{9.5}$$

The admittance of a series resonant circuit can be expressed in terms of the generalized detuning as

$$Y = \frac{1/R}{1 + j\xi} = |Y| \exp(j\varphi) \tag{9.6}$$

From Eq. (9.6), the amplitude response of the resonant circuit can be written as

Solve Problem 1

$$|Y| = \frac{1/R}{\sqrt{1+\xi^2}} \tag{9.7}$$

and the phase response as

$$\varphi = -\arctan \xi \tag{9.8}$$

The plots constructed on the basis of Eqs. (9.7) and (9.8) appear in Fig. 9.1.

The magnitude of the admittance decreases to $1/\sqrt{2}$, or 0.707 times the resonant value at frequencies such that $\xi = \pm 1$. Also,

$$Q\left(\frac{\omega_{res} + \Delta\omega}{\omega_{res}} - \frac{\omega_{res}}{\omega_{res} + \Delta\omega}\right) = \pm 1$$

If $Q \gg 1$, the equality will be satisfied at low values of the ratio $\Delta\omega/\omega_{res}$. Therefore, approximately

$$2Q\Delta\omega/\omega_{res} = \pm 1$$

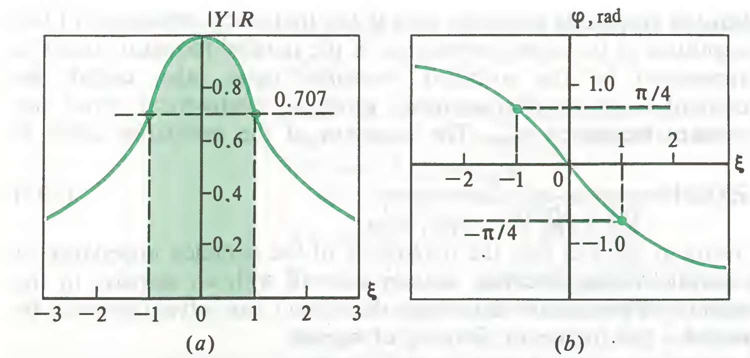
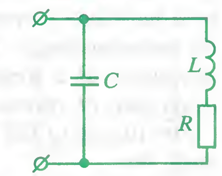


Fig. 9.1 Admittance of a series resonant circuit as a function of generalized detuning: (*a*) magnitude; (*b*) phase

Hence, the bandwidth of the resonant circuit between the 0.707 points will be

$$BW_{0.707} = 2|\Delta\omega| = \omega_{res}/Q \tag{9.9}$$



Frequency response of a parallel resonant circuit. At the resonance frequency a series resonant circuit has a very low impedance the magnitude of which sharply increases with the amount off resonance (detuning). For a parallel resonant circuit the impedance at resonance

$$R_{res} = \rho Q \tag{9.10}$$

being likewise purely resistive, is a maximum in magnitude.

Taking the input as a current, and the output as a voltage, the frequency response of a parallel resonant circuit is called the *complex input (driving-point) impedance*, $Z(j\omega)$. It is convenient to express the frequency response function, taking the generalized detuning ξ as the argument:

$$Z(j\xi) = \frac{R_{res}}{1 + j\xi} \tag{9.11}$$

The plots of the amplitude and phase of $Z(j\xi)$ are exactly the same as appear in Fig. 9.1 for the complex admittance of a series resonant circuit.

If the Q-factor is sufficiently large so that in the direct vicinity of the resonant frequency we may express the generalized detuning by an approximate formula of the form

$$\xi \approx \frac{2Q(\omega - \omega_{res})}{\omega_{res}} \tag{9.12}$$

Solve Problem 2

then the amplitude response, that is the frequency depedence of the magnitude of the input impedance, of the parallel resonant circuit is represented by the so-called *resonance curve* (also called the resonant curve or characteristic) which is symmetrical about the resonant frequency ω_{res}. The equation of the resonance curve is

$$|Z(j\omega)| = \frac{R_{res}}{\sqrt{1+4Q^2(\omega-\omega_{res})^2/\omega_{res}^2}} \tag{9.13}$$

Owing to the fact that the magnitude of the complex impedance of a parallel resonant circuit sharply falls off with an increase in the amount off resonance (detuning), this circuit can advantageously be used for the frequency filtering of signals.

Example 9.1. *A parallel resonant circuit for which $Q = 125$ and $L = 6$ μH is tuned to resonate at $f_{res} = 8$ MHz. The circuit is driven by a harmonic current source; the output signal is the voltage across the resonant circuit. Determine the amount by which the signal will be attenuated at a frequency of* 8.1 MHz *as compared with the signal magnitude at resonance.*

For tuning to the desired resonant frequency, the capacitor rating must be

$$C = 1/4\pi^2 Lf_{res}^2 = 66 \text{ pF}$$

The resonant impedance of the circuit is

$$R_{res} = \rho Q = \sqrt{L/C}\,Q = 37.69 \text{ k}\Omega$$

By virtue of Eq. (9.12), the generalized detuning at 8.1 MHz is

$$\xi = 2Q\Delta f/f_{res} = 3.125$$

Now the amplitude of the output signal is proportional to the

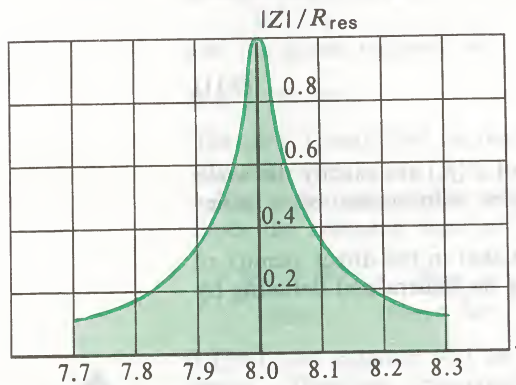


Fig. 9.2 Amplitude response of a parallel resonant circuit with $Q = 125$, $L = 6$ μH, and $C = 66$ pF

magnitude of the resonant-circuit impedance. Since

$$|Z|/R_{res} = 1/\sqrt{1+\xi^2}$$

then, on substituting the found value of ξ, we find that the signal amplitude at frequency 8.1 MHz is 0.305 times the amplitude at the resonant frequency.

This corresponds to a negative gain (loss or attenuation)

$$\Delta = 20\log_{10} 0.305 = -10.31 \text{ dB}$$

The amplitude response of the system is shown in Fig. 9.2.

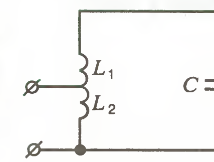
The specific shape of the amplitude response is an indication that the resonant circuit in question is a narrowband frequency-selective system. This is true because the ratio of the bandwidth to the resonant frequency is

$$BW_{0.707}/f_{res} = 1/Q = 8 \times 10^{-3} \ll 1$$

Frequently, use is made of tapped parallel resonant circuits. This means that the associated external circuit is connected to a tap down on the tuned-circuit coil (inductor). The input impedance of a tapped resonant circuit is found by Eq. (9.11) in which one inserts the purely resistive resonant impedance

$$R_{res} = k_{t.d}^2 \rho Q$$

where $k_{t.d} = L_2/(L_1+L_2)$ is the tapping-down factor with the inductive coupling neglected.



Connection across part of the tuned-circuit inductor instead of the whole of it brings down the resonant impedance without extending the bandwidth

The pole-zero representation of the response of resonant circuits. Consider Eq. (9.11) for the complex input impedance of a parallel resonant circuit and express the generalized detuning ξ in terms of the current frequency. Then

$$Z(j\omega) = \frac{R_{res}}{1+jQ(\omega/\omega_{res}-\omega_{res}/\omega)} \tag{9.14}$$

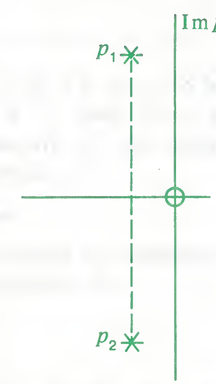
On changing from the frequency variable $j\omega$ to the complex frequency p, we obtain the Laplace transform of the tuned-circuit impedance

$$Z(p) = \frac{R_{res}}{1+Q(p/\omega_{res}+\omega_{res}/p)} = \frac{p/C}{p^2+p\omega_{res}/Q+\omega_{res}^2}$$

The transform $Z(p)$ has an only zero for $p = 0$ and two complex conjugate roots

$$p_{1,2} = -\omega_{res}/2Q \pm j\sqrt{1-1/4Q^2}\,\omega_{res} \tag{9.15}$$

The poles are located in the left-hand half-plane (the system is stable), and as their distance to the imaginary axis decreases the Q-factor of the system improves. This property is common to all frequency-selective systems.



The pole Q-factor

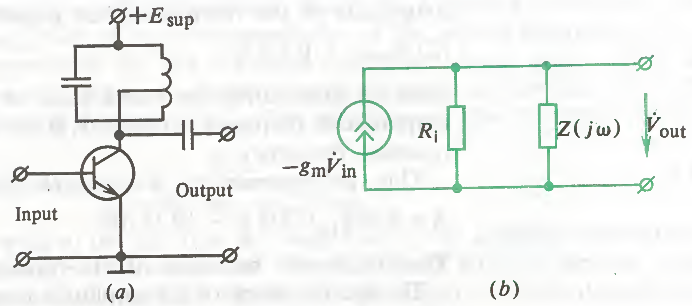


Fig. 9.3 Small-signal tuned amplifier: (a) schematic diagram; (b) equivalent circuit

Sometimes it is convenient to introduce a special numeric characteristic called the *pole Q*, or Q_{pole}, defined as the ratio of the absolute values of the imaginary and real parts of the pole coordinates. As follows from Eq. (9.15), in our case

$$Q_{pole} = |\mathrm{Im}\, p_{1,2}|/|\mathrm{Re}\, p_{1,2}| \approx 2Q$$

The small-signal tuned amplifier. This system combines the functions of an amplifier and of a linear frequency filter (Fig. 9.3). It differs from an *RC*-loaded amplifier (see Chap. 8) in that the load is a parallel resonant circuit. Generally, the electron device of the amplifier (which may be an electron tube or a transistor) is tapped down on, that is, connected across part of, the tuned-circuit coil.

Referring to the equivalent circuit of the amplifier, it is seen that the current $-g_m\dot{V}_{in}$ supplied by the controlled source flows through an impedance given by

$$Z_{eq}(j\omega) = \frac{Z(j\omega)R_i}{Z(j\omega)+R_i}$$

and produces across it a voltage drop which is the output signal of the amplifier. Simple manipulations show [see Eq. (9.11)] that

$$Z_{eq}(j\omega) = \frac{R_{res,\,eq}}{1 + j\xi_{eq}} \tag{9.16}$$

where

$$R_{res,\,eq} = \frac{R_{res}}{1 + R_{res}/R_i} \tag{9.17}$$

is the equivalent impedance of the amplifier tuned circuit at resonance, as corrected for the internal resistance of the source

$$\xi_{eq} = \frac{\xi}{1 + R_{res}/R_i}$$

It may be taken that the effect of the internal resistance of the source consists in bringing down the Q-factor of the resonant system so that it becomes equal to an equivalent Q-factor

$$Q_{eq} = \frac{Q}{1 + R_{res}/R_i} \tag{9.18}$$

As follows from Eq. (9.18), in order to minimize the shunting effect of the electron device on the tuned circuit without having to expand in the bandwidth of the amplifier, one should bring down the resonant impedance R_{res} by tapping the electron device down on the tuned-circuit coil. Since the complex amplitude of the harmonic signal at the amplifier output is

$$\dot{V}_{out} = -g_m Z_{eq}\dot{V}_{in}$$

the frequency response of the system is

Work Problem 5

$$K(j\xi_{eq}) = -g_m R_{res,\,eq}/(1 + j\xi_{eq}) \tag{9.19}$$

Hence, the amplitude response and the phase response of the amplifier have the form

$$|K(j\omega)| = \frac{g_m R_{res,\,eq}}{\sqrt{1 + \dfrac{4Q_{eq}^2(\omega - \omega_{res})^2}{\omega_{res}^2}}} \tag{9.20}$$

$$\varphi_K(\omega) = \pi - \arctan\frac{2Q_{eq}(\omega - \omega_{res})}{\omega_{res}} \tag{9.21}$$

Example 9.2. *For the amplifier shown in Fig. 9.3,* $f_{res} = 28$ MHz, $Q = 95$, $\rho = 430\ \Omega$, $k_{t.d} = 0.6$, $g_m = 20$ mA V^{-1}, and $R_i = 15$ kΩ. *Find the gain at resonance and the bandwidth of the amplifier.*

The resonant impedance of the tuned circuit is

$$R_{res} = k_{t.d}^2\rho Q = 0.36 \times 0.43 \times 95 = 14.17\ \text{k}\Omega$$

The equivalent impedance of the tuned circuit at resonance, as corrected for the shunting effect of the transistor, is

$$R_{res,\,eq} = \frac{14.71}{1 + 14.71/15} = 7.43\ \text{k}\Omega$$

At resonance, $\xi_{eq} = 0$, so, as follows from (9.19), the gain at resonance is

$$K_{res} = g_m R_{res,\,eq} = 148.6$$

or in decibels

$\Delta_{res} = 20 \log_{10} K_{res} = 43.44$ dB

The bandwidth of the amplifier between the 0.707 points is found by Eq. (9.9)

$\mathrm{BW}_{0.707} = f_{res}/Q_{eq} = 0.584$ MHz

Multistage frequency-selective systems. The simple narrowband circuits we have examined suffer from a serious drawback – a low frequency selectivity. This property manifests itself in that outside the passband their amplitude response rolls off slowly or, which is the same, the rate of cut-off is low. Therefore, the output wave contains not only the valid signal whose spectrum lies near the peak of the amplitude response, but also a definite, sometimes considerable proportion of interfering signals, noise, etc., with their spectra lying rather far from the frequency to which the filter is tuned.

One way to improve the frequency selectivity of filters is to use several resonant circuits in tandem, so that the overall amplitude response is nearly rectangular.

A simple example of multistage frequency-selective systems is a system of two inductively coupled tuned circuits. Its operating principle is examined in a course on circuit theory. Figure 9.4*a* shows the schematic diagram of a tuned amplifier loaded into a system of two identical inductively coupled tuned circuits. The important parameters of this system are the coupling coefficient (or coefficient of coupling), $k_c = M/L$, and the coupling

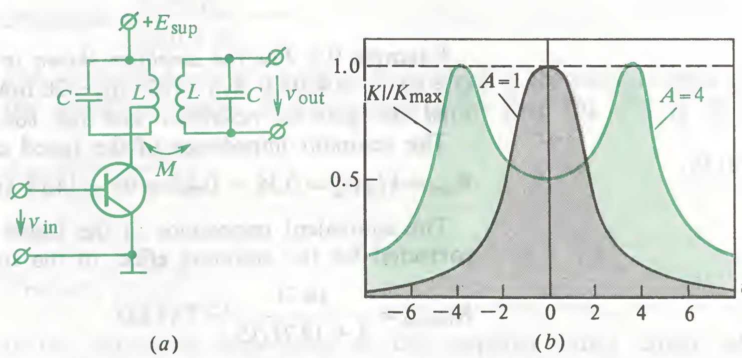


Fig. 9.4 Inductively-coupled tuned amplifier: (*a*) schematic diagram; (*b*) plots of amplitude response for several values of the coupling factor

factor, $A = k_c Q$. The amplitude response of this amplifier is defined as

$$|K(j\xi)| = \frac{k_{t.d} A g_m R_{res,eq}}{\sqrt{(1 + A^2 - \xi^2)^2 + 4\xi^2}} \qquad (9.22)$$

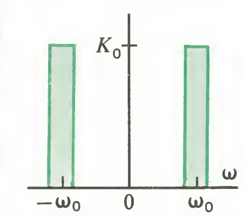
Plots of the amplitude response based on (9.22) are constructed for several values of A in Fig. 9.4*b*. It is to be noted that if $A > 1$, the resonance curve within the pass band has a dip which increases with increasing A. If we compare the amplitude responses of a single-tuned and a double-tuned amplifier, we will note that, given the same Q-factor, the double-tuned-circuit amplifier has a resonance curve with a faster rate of roll-off, that is, a better frequency selectivity.

By using a large number of inductively coupled resonant circuits connected in tandem, it is possible to synthesize highly efficient frequency-selective systems.

A recent trend in communication practice has been to use active frequency-selective filters based on novel circuit-engineering principles (see Chap. 14). Much headway has been made in the design of filters utilizing wave phenomena in solids. This new field, called acoustoelectronics, holds out great promise for the synthesis of miniature and reliable frequency-selective systems.

Idealized models of frequency-selective devices. When treating frequency-selective systems theoretically, it is advantageous to use their simplified models which adequately represent the most important aspects of filter behaviour and omit minor details which are hard to analyse anyway.

The ideal bandpass filter

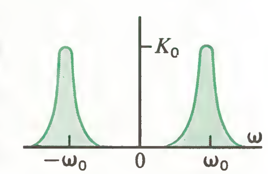


The simplest model in this class is the hypothetical *ideal bandpass filter* which has a flat frequency response equal to K_0 within the pass band:

$$K(j\omega) = \begin{cases} K_0 & \text{for } -\omega_0 - \Delta\omega < \omega < -\omega_0 + \Delta\omega \\ K_0 & \text{for } \omega_0 - \Delta\omega < \omega < \omega_0 + \Delta\omega \\ 0 & \text{elsewhere} \end{cases} \qquad (9.23)$$

Another widely used theoretical model of narrowband systems is the so-called *Gaussian radio filter* for which the amplitude response function is a bell-shaped Gaussian curve symmetrical about frequency ω_0. The frequency response of a Gaussian radio filter is

The Gaussian radio filter



$$K(j\omega) = K_0 \exp[-b(\omega + \omega_0)^2] + K_0 \exp[-b(\omega - \omega_0)^2] \qquad (9.24)$$

Here, b is a constant defining the frequency properties of the filter. The first term in (9.24) represents a "spike" in the negative-frequency region, and the second term, a "spike" in the positive-frequency region. If $b\omega_0^2 \gg 1$, the filter is a narrowband one, and no overlap occurs between the frequency responses corresponding to the negative and positive frequencies.

9.2 Response of Frequency-Selective Circuits to Broadband Excitations

Solve Problem 6

It is interesting to know the response of a narrowband frequency-selective circuit to a broadband excitation because interfering signals are frequently very short pulses. Since the bandwidth of a pulse increases as its duration is decreased, the effective bandwidth of such interfering signals may substantially exceed the pass band of the frequency-selective system involved.

The concept of broadband signal. Let $K(j\omega)$ be the frequency response of a frequency-selective circuit capable of extracting the spectral components of the input situated in the vicinity of frequencies $\pm\omega_0$. The excitation $u_{in}(t)$ with spectrum $S_{in}(\omega)$ is called *broadband* with regard to the given circuit, if $S_{in}(\omega)$ may be taken as being approximately constant within the pass band of the system. Then,

$$u_{out} \approx \frac{S_{in}(-\omega_0)}{2\pi} \int_{-\infty}^{0} K(j\omega)\exp(j\omega t)\,d\omega + \frac{S_{in}(\omega_0)}{2\pi} \int_{0}^{\infty} K(j\omega)\exp(j\omega t)\,d\omega \qquad (9.25)$$

As follows from Eq. (9.25), *the waveform of the output signal is decided by the frequency response of the system, and not by the shape of the excitation.* The spectrum of the excitation within the pass band of the system determines the scale of the response.

The impulse response of a frequency-selective circuit. A signal with an infinitely broad bandwidth is the delta impulse for which $S_{in}(\omega) = 1$. In this case, the output signal is the impulse response $h(t)$. According to (9.25),

$$h(t) = \frac{1}{2\pi}\int_{-\infty}^{0} K(j\omega)\exp(j\omega t)\,d\omega + \frac{1}{2\pi}\int_{0}^{\infty} K(j\omega)\exp(j\omega t)\,d\omega \qquad (9.26)$$

Consider the first term on the right-hand side of (9.26) and replace its integration variable ω with a new frequency variable Ω:

$$\omega = -\omega_0 - \Omega$$

This change implies that the frequency response function $K(j\omega)$ is moved from the neighbourhood of the frequency $-\omega_0$ into the neighbourhood of the point $\Omega = 0$. Therefore,

$$\frac{1}{2\pi}\int_{-\infty}^{0} K(j\omega)\exp(j\omega t)\,d\omega = \frac{-\exp(-j\omega_0 t)}{2\pi} \times \int_{\infty}^{-\omega_0} K[-j(\omega_0+\Omega)]\exp(-j\Omega t)\,d\Omega$$

$$= \frac{\exp(-j\omega_0 t)}{2\pi}\int_{-\omega_0}^{\infty} K[-j(\omega_0+\Omega)]\exp(-j\Omega t)\,d\Omega \qquad (9.27)$$

Since the circuit in question is a narrowband one, the amplitude response sharply decreases with increasing Ω. This means that in the last integral of (9.27) the lower limit $-\omega_0$ may be replaced with $-\infty$:

$$\frac{1}{2\pi}\int_{-\infty}^{0} K(j\omega)\exp(j\omega t)\,d\omega = \frac{\exp(-j\omega_0 t)}{2\pi} \times \int_{-\infty}^{\infty} K[-j(\omega_0+\Omega)]\exp(-j\Omega t)\,d\Omega \qquad (9.28)$$

Similarly, by a change of variable, $\omega = \omega_0 + \Omega$, the second integral in (9.26) can be re-cast as

$$\frac{1}{2\pi}\int_{0}^{\infty} K(j\omega)\exp(j\omega t)\,d\omega = \frac{\exp(j\omega_0 t)}{2\pi}\int_{0}^{\infty} K[j(\omega_0+\Omega)]\exp(j\Omega t)\,d\Omega \qquad (9.29)$$

Since the complex conjugate expressions (9.28) and (9.29) combine, the expression for the impulse response of a narrowband system takes the form

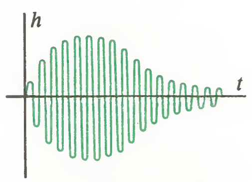
$$h(t) = 2\mathrm{Re}\left\{\frac{1}{2\pi}\int_{-\infty}^{\infty} K[j(\omega_0+\Omega)]\exp(j\Omega t)\,d\Omega\exp(j\omega_0 t)\right\} \qquad (9.30)$$

The low-frequency equivalent of a frequency-selective network. This refers to an imaginary system whose frequency response is obtained by shifting that of a real narrowband network from the vicinity of frequency ω_0 into the neighbourhood of zero frequency, that is,

$$K_{lf}(j\Omega) = K[j(\omega_0+\Omega)] \qquad (9.31)$$

The integral in (9.30) is the impulse response of the l.f. equivalent

$$h_{lf}(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} K_{lf}(j\Omega)\exp(j\Omega t)\,d\Omega \qquad (9.32)$$



The typical impulse response of a frequency-selective network

Therefore,

$$h(t) = \mathrm{Re}\left[2h_{\mathrm{lf}}(t)\exp(j\omega_0 t)\right] \tag{9.33}$$

It follows then that the function $2h_{\mathrm{lf}}(t)$ is the complex envelope of the impulse response of a real narrowband network. As follows from (9.33), the impulse response of a frequency-selective system is, in the general case, a quasiharmonic wave the envelope and initial phase of which vary slowly in time (on the time scale $T = 2\pi/\omega_0$).

Example 9.3. *The l.f. equivalent of a parallel resonant circuit.*

Here, the frequency response of the system is its complex input impedance

$$Z(j\omega) = \frac{R_{\mathrm{res}}}{1 + j\dfrac{2Q(\omega - \omega_{\mathrm{res}})}{\omega_{\mathrm{res}}}} \tag{9.34}$$

The frequency response of the l.f. equivalent can be obtained by a change of variable $\omega = \omega_{\mathrm{res}} + \Omega$ in (9.34):

$$Z(j\Omega) = \frac{R_{\mathrm{res}}}{1 + j\dfrac{2Q\Omega}{\omega_{\mathrm{res}}}} \tag{9.35}$$

To within the scale factor R_{res}, this is the frequency response of a 1st-order dynamic system (similar to an RC-network) with a time constant

The tuned-circuit time constant

$$\tau_{\mathrm{ckt}} = 2Q/\omega_{\mathrm{res}} \tag{9.36}$$

called the *tuned-circuit time constant.*

The impulse response of such a network has been found in Chap. 8 when examining the properties of the RC-network:

$$h_{\mathrm{lf}}(t) = (R_{\mathrm{res}}/\tau_{\mathrm{ckt}})\exp(-t/\tau_{\mathrm{ckt}}) \quad \text{for } t > 0 \tag{9.37}$$

Thus, the impulse response of a parallel resonant circuit is

$$h(t) = (R_{\mathrm{res}}\omega_{\mathrm{res}}/Q)\exp(-t/\tau_{\mathrm{ckt}})\cos\omega_{\mathrm{res}} t \tag{9.38}$$

Since

$$R_{\mathrm{res}}\omega_{\mathrm{res}}/Q = 1/C$$

where C is the resonant-circuit capacitance, the result thus obtained fully checks with that found in Example 8.20.

Example 9.4. *Find the impulse response of an idealized narrowband system whose frequency response is*

$$K(j\omega) = K_0 \exp[-b(\omega - \omega_0)]\,\sigma(\omega - \omega_0)$$

for $\omega > 0$.

On shifting the above function into the vicinity of zero frequency, we obtain the frequency response of the l.f. equivalent

$$K_{\mathrm{lf}}(j\Omega) = K_0 \exp(-b\Omega)\,\sigma(\Omega) \tag{9.39}$$

Hence, the corresponding impulse response is

$$h_{\mathrm{lf}}(t) = \frac{K_0}{2\pi}\int_0^{\infty} \exp[-(b - jt)\Omega]\,d\Omega = K_0/2\pi(b - jt) \tag{9.40}$$

It is to be noted that $h_{\mathrm{lf}}(t)$ is a complex-valued function. Therefore, the l.f. equivalent of the filter in question is not a physically realizable network. This, however, does not stand in the way of obtaining the impulse response of the original system by use of Eq. (9.33):

$$h(t) = \mathrm{Re}\left[\frac{K_0}{\pi(b - jt)}\exp(j\omega_0 t)\right] = \frac{K_0}{\pi}\frac{1}{b^2 + t^2}(b\cos\omega_0 t - t\sin\omega_0 t)$$

Example 9.5. *Find the impulse response of the double-stage tuned amplifier whose schematic diagram is shown in Fig. 9.5.*

Assume for simplicity that the two stages are tuned to the same resonant frequency ω_{res}, have the same gain at resonance K_{res}, and the same tuned-circuit time constant τ_{ckt}. Then the frequency response of the amplifier is

$$K(j\omega) = \frac{K_{\mathrm{res}}^2}{[1 + j\tau_{\mathrm{ckt}}(\omega - \omega_{\mathrm{res}})]^2}$$

Hence,

$$K_{\mathrm{lf}}(j\Omega) = K_{\mathrm{res}}^2/(1 + j\Omega\tau_{\mathrm{ckt}})^2$$

On replacing the frequency variable $j\Omega$ with the complex fre-

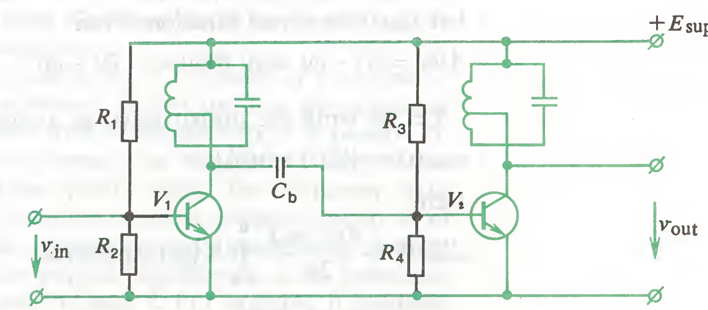


Fig. 9.5 Two-stage amplifier (the circuit components printed in black perform auxiliary functions): R_1 through R_4 apply initial bias to transistors T_1 through T_4; C_b–d.c. blocking capacitor

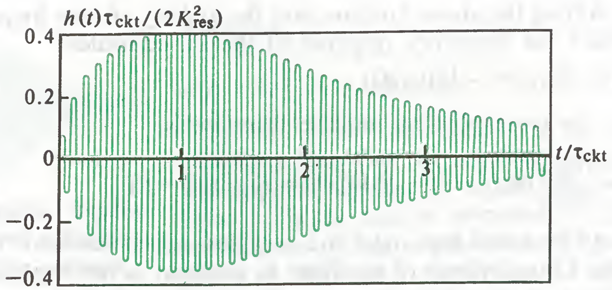


Fig. 9.6 Impulse response of a two-stage tuned amplifier

quency p, we get

$$K_{lf}(p) = K_{res}^2/(1+p\tau_{ckt})^2 = (K_{res}^2/\tau_{ckt}^2)/(p+1/\tau_{ckt})^2$$

From tables of Laplace transform pairs we find that the above transfer function corresponds to the impulse response

$$h_{lf}(t) = (K_{res}^2/\tau_{ckt}^2)\, t \exp(-t/\tau_{ckt})$$

Hence, the impulse response of the two-stage amplifier is

$$h(t) = (2K_{res}^2/\tau_{ckt})(t/\tau_{ckt})\exp(-t/\tau_{ckt})\cos\omega_{res}t \tag{9.41}$$

A plot of the impulse response is given in Fig. 9.6.

It is instructive to compare this result with the impulse response of the single parallel tuned circuit in Example 8.20. As is seen, the output pulse is "pulled" in time because of the appreciable time lag associated with a double-tuned-circuit system.

The general case. Suppose that a frequency-selective system is driven by an arbitrary broadband signal whose spectrum is

$$S_{in}(\omega) = A(\omega) + jB(\omega)$$

Let $u_{in}(t)$ be a real function. Then

$$A(\omega) = A(-\omega) \text{ and } B(\omega) = -B(-\omega)$$

Let us write the output wave as a sum:

$$u_{out}(t) = u_{out}^{(1)}(t) + ju_{out}^{(2)}(t) \tag{9.42}$$

Here,

$$u_{out}^{(1)}(t) = \frac{A(-\omega_0)}{2\pi}\int_{-\infty}^{0} K(j\omega)\exp(j\omega t)\,d\omega + \frac{A(\omega_0)}{2\pi}\int_{0}^{\infty} K(j\omega)\exp(j\omega t)\,d\omega$$

$$= 2A(\omega_0)\operatorname{Re}[h_{lf}(t)\exp(j\omega_0 t)] \tag{9.43}$$

Similarly,

$$u_{out}^{(2)}(t) = \frac{B(-\omega_0)}{2\pi}\int_{-\infty}^{0} K(j\omega)\exp(j\omega t)\,d\omega + \frac{B(\omega_0)}{2\pi}\int_{0}^{\infty} K(j\omega)\exp(j\omega t)\,d\omega$$

$$= j2B(\omega_0)\operatorname{Im}[h_{lf}(t)\exp(j\omega_0 t)] \tag{9.44}$$

Work Problem 7

Substituting (9.43) and (9.44) in (9.42) finally yields

$$u_{out}(t) = 2\{A(\omega_0)\operatorname{Re}[h_{lf}(t)\exp(j\omega_0 t)] - B(\omega_0)\operatorname{Im}[h_{lf}(t)\exp(j\omega_0 t)]\} \tag{9.45}$$

As a special case, Eq. (9.33) describing the impulse response of a narrowband network naturally stems from (9.45).

The physical significance of the spectral expansion. Assume for simplicity that $h_{lf}(t)$ is a real function, and write Eq. (9.45) as

$$u_{out}(t) = 2h_{lf}(t)[A(\omega_0)\cos\omega_0 t - B(\omega_0)\sin\omega_0 t]$$

$$= 2h_{lf}(t)\sqrt{A^2+B^2}\cos(\omega_0 t + \varphi) \tag{9.46}$$

where $\varphi = \arctan(B/A)$ and $\sqrt{A^2+B^2} = |S_{in}(\omega_0)|$.

Thus, we may conclude that the response of a narrowband network to a broadband signal is proportional to the absolute value of the spectrum of the excitation at the point on the frequency axis corresponding to the central frequency of the network's pass band.

This suggests an approach to the hardware implementation of signal spectrum analysis. Figure 9.7 shows the block diagram of the parallel type of *spectrum analyzer*.

The spectrum analyzer shown in Fig. 9.7 consists of a number of narrowband filters whose passbands do not overlap. By measuring the amplitudes of the output waves simultaneously, it is possible to obtain information about the intensity of the spectral components of the input signal at some points along the frequency axis. The above principle of spectrum analysis implementation is of important significance both practically and theoretically. Among other things, it brings out the physical significance of the behaviour of the signal spectra examined in Chap. 2. For example, it has been found that the spectrum of a rectangular video pulse of duration τ_0 is zero at all frequencies $\omega_n = 2\pi n/\tau_p$ $(n = 1, 2, \ldots)$. Suppose that this video pulse drives a very narrowband resonant network tuned to one of these frequencies. The ratio of the natural period of the

Fig. 9.7 Block diagram of a signal spectrum analyser. (The amplitudes of the filter outputs are proportional to the magnitudes of the spectra.)

network to the pulse duration: $T_n = 2\pi/\omega_n = \tau_p/n$ is an integer. On being driven by the leading edge of the input pulse in one direction, the system will be "kicked" by the trailing edge of the same pulse in the opposite direction in a time which is a multiple of the network's natural period. As a result, the two responses will cancel each other. Precisely this occurrence is corroborated by the zero values of the spectrum of the video pulse at some points along the frequency axis.

An excellent presentation of material on the physical aspects of spectral expansion will be found in [7].

9.3 Response of Frequency-Selective Circuits to Narrowband Excitations

In a typical situation, such as in the reception of modulated signals, a frequency-selective linear filter is driven by a valid signal the spectrum of which has a well-defined peak within the pass band of the filter. As a rule, the resonant frequency of the tuned circuit is the same as the carrier frequency (symmetric tuning).

The cause of signal distortion in frequency-selective systems

If the spectrum of the input signal were strictly bounded to the frequency band within which the frequency response of the filter remains constant, the output signal would be a scaled replica of the input signal. However, both the amplitude response and the phase response of a frequency-selective network are far from ideal, and this leads to distortion in the waveform of the output signal. A method by which one can determine signals at the output of frequency selective systems driven by narrowband waves is set forth below.

Basic relations. Let us consider an arbitrary narrowband network whose frequency response $K(j\omega)$ is substantially nonzero only in the vicinity of points $\pm\omega_0$ on the frequency axis. Suppose that the excitation is a narrowband (quasiharmonic) wave whose central frequency is ω_0. This means that in the equation

$$u_{in}(t) = \mathrm{Re}[\tilde{U}_{in}(t)\exp(j\omega_0 t)] \tag{9.47}$$

the complex envelope $\tilde{U}_{in}(t)$ varies more slowly than the wave $\cos\omega_0 t$. Let the correspondence between the signals and their spectra be designated as

$$u_{in}(t) \leftrightarrow S_{in}(\omega), \quad \tilde{U}_{in}(t) \leftrightarrow G_{in}(\omega)$$

such that (see Chap. 5) the spectra of the excitation and of its envelope are connected by a relation of the form:

$$S_{in}(\omega) = \frac{1}{2}G_{in}(\omega - \omega_0) + \frac{1}{2}G^*_{in}(-\omega - \omega_0)$$

Hence, by using the fundamental equation of the spectral method, we obtain the following expression for the output signal:

$$u_{out}(t) = \frac{1}{4\pi}\int_{-\infty}^{0} G^*_{in}(-\omega - \omega_0)K(j\omega)\exp(j\omega t)\,d\omega$$

$$+ \frac{1}{4\pi}\int_{0}^{\infty} G_{in}(\omega - \omega_0)K(j\omega)\exp(j\omega t)\,d\omega \tag{9.48}$$

By a change of variable, $\omega = -\omega_0 - \Omega$, in the first integral, we may re-write it as

$$\frac{1}{4\pi}\int_{-\infty}^{0} G^*_{in}(-\omega - \omega_0)K(j\omega)\exp(j\omega t)\,d\omega$$

$$= \frac{1}{4\pi}\int_{-\infty}^{\infty} G^*_{in}(\Omega)K[-j(\omega_0 + \Omega)]\exp(-j\Omega t)\,d\Omega\exp(-j\omega_0 t) \tag{9.49}$$

Similarly, on inserting $\omega = \omega_0 + \Omega$, the second integral in (9.48) may be re-written as

$$\frac{1}{4\pi}\int_{0}^{\infty} G_{in}(\omega - \omega_0)K(j\omega)\exp(j\omega t)\,d\omega$$

$$= \frac{1}{4\pi}\int_{-\infty}^{\infty} G_{in}(\Omega)K[j(\omega_0 + \Omega)]\exp(j\Omega t)\,d\Omega\exp(j\omega_0 t) \tag{9.50}$$

In taking the sum of the right-hand sides of (9.49) and (9.50), it is to be noted that the two expressions are complex conjugates of each other. Also, on the basis of Eq. (9.31), the term $K[j(\omega_0 + \Omega)]$ is the frequency response of the l.f. equivalent of a narrowband network. Therefore,

$$u_{out}(t) = \mathrm{Re}\left[\frac{1}{2\pi}\int_{-\infty}^{\infty} G_{in}(\Omega)K_{lf}(j\Omega)\exp(j\Omega t)\,d\Omega\exp(j\omega_0 t)\right]$$

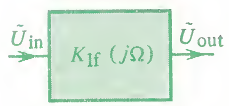
Hence, the expression for the complex envelope of the output signal is

$$\tilde{U}_{out}(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} G_{in}(\Omega) K_{lf}(j\Omega) \exp(j\Omega t)\, d\Omega \tag{9.51}$$

It should be interpreted as follows: *The complex envelope of the output signal is a slowly varying wave whose spectrum is*

$$G_{out}(\Omega) = G_{in}(\Omega) K_{lf}(j\Omega) \tag{9.52}$$

In order to find the output signal of a frequency-selective network driven by a narrowband signal, we should first determine the response of the l.f. equivalent of the original network to the complex envelope of the excitation. The final step will then be to recover the physical output signal:



$$u_{out}(t) = \mathrm{Re}[\tilde{U}_{out}(t) \exp(j\omega_0 t)] \tag{9.53}$$

Equation (9.51) corresponds to the spectral (frequency-domain) method of finding the output signal of an l.f. equivalent. Other methods, such as the Laplace transformation and the Duhamel superposition integral, may be used as well. In accord with the Duhamel superposition integral method,

$$\tilde{U}_{out}(t) = \int_{-\infty}^{t} \tilde{U}_{in}(\tau) h_{lf}(t-\tau)\, d\tau \tag{9.54}$$

where $h_{lf}(t)$ is the impulse response of the l.f. equivalent.

Response of a single-stage tuned amplifier to an AM signal. We seek to find the response to a single-tone AM wave

$$u_{in}(t) = U_0(1 + M\cos\Omega t)\cos\omega_0 t \tag{9.55}$$

of a single-stage tuned amplifier whose frequency response is

$$K(j\omega) = \frac{-K_{res}}{1 + j\tau_{ckt}(\omega - \omega_{res})} \tag{9.56}$$

Let us adopt an important simplifying assumption that the resonant frequency ω_{res} and the carrier frequency ω_0 are the same. The complex envelope of the input signal is

$$\tilde{U}_{in}(t) = U_0(1 + M\cos\Omega t) \tag{9.57}$$

The frequency response of the l.f. equivalent of the amplifier is

$$K(j\Omega) = -K_{res}/(1 + j\Omega\tau_{ckt}) \tag{9.58}$$

The complex envelope of the output signal may be found from (9.57) and (9.58) by the usual method of complex amplitudes known from circuit theory:

$$\tilde{U}_{out}(t) = -K_{res}U_0 - \frac{K_{res}U_0 M}{\sqrt{1+\Omega^2\tau_{ckt}^2}}\cos(\Omega t - \vartheta)$$

where $\vartheta = \arctan \Omega\tau_{ckt}$. On substituting this expression in (9.53), we finally obtain

$$u_{out}(t) = -K_{res}U_0\left[1 + \frac{M}{\sqrt{1+\Omega^2\tau_{ckt}^2}}\cos(\Omega t - \vartheta)\right]\cos\omega_0 t$$

Since the tuned-circuit time constant is

$$\tau_{ckt} = 2Q_{res}/\omega_{res}$$

the above relation may be re-cast as

$$u_{out}(t) = -K_{res}U_0\left[1 + \frac{M}{\sqrt{1+\xi_\Omega^2}}\cos(\Omega t - \vartheta)\right]\cos\omega_0 t \tag{9.59}$$

where $\xi_\Omega = 2Q_{res}\Omega/\omega_{res}$ is the generalized detuning of the amplifier tuned circuit at the upper side frequency.

Thus, the wave appearing at the amplifier output is likewise an AM signal, but boosted in amplitude. It differs from the input signal in that it has a smaller modulation factor (or modulation depth):

Solve Problem 8

$$M_{out} = M_{in}/\sqrt{1+\xi_\Omega^2} \tag{9.60}$$

Also, the envelope of the output signal is delayed from the envelope of the input signal by a time $t_d = \vartheta_\Omega/\Omega$.

Example 9.6. *Let an AM-signal of $M = 0.8$, $\omega_0 = 5 \times 10^6\ s^{-1}$ and $\Omega = 3 \times 10^4\ s^{-1}$ drive an amplifier tuned to the carrier frequency. The equivalent Q-factor of the amplifier is $Q_{eq} = 75$.*

In this case,

$$\xi_\Omega = 2 \times 75 \times 3 \times 10^4/5 \times 10^6 = 0.9$$

Hence, by virtue of Eq. (9.60),

$$M_{out} = 0.8/\sqrt{1+0.81} = 0.595$$

As is seen, the depth of modulation is substantially reduced. Because $\arctan 0.9 = 0.733$ rad, the envelope is delayed for a time $t_d = 0.733/(3 \times 10^4) = 24.43\ \mu s$.

Response to a truncated voltage sinewave. In many

communication systems (radar, multichannel radio links, etc.), the useful information is transmitted by means of sequences of rectangular radio pulses. On passing through resonant frequency-selective networks which are integral parts of the respective receivers, these input pulses are somewhat distorted. In order to assess the degree of distortion, let us determine the output signal of a single-stage tuned amplifier whose frequency response is given by (9.56), when the input signal is

$$v_{in}(t) = V_m \cos \omega_0 t\sigma(t)$$

If $\omega_{res} = \omega_0$, then, on taking this frequency as reference, we obtain the following expression for the complex envelope:

$$\tilde{V}_{in}(t) = V_m \sigma(t) \tag{9.61}$$

The effect of the signal defined in (9.61) on a system whose frequency response has the form of (9.58) has been examined in Chap. 8 in connection with the step response of an RC-network. Therefore, by taking advantage of the already known result, we may write

$$\tilde{V}_{out}(t) = -K_{res} V_m [1 - \exp(-t/\tau_{ckt})]\sigma(t) \tag{9.62}$$

Hence, the output signal of the amplifier is

▲ Solve Problem 9

$$v_{out}(t) = -K_{res} V_m [1 - \exp(-t/\tau_{ckt})] \cos \omega_0 t \tag{9.63}$$

The plot corresponding to (9.63) appears in Fig. 9.8.

● **The rise time of an oscillatory system**

The instantaneous amplitude of the input signal reaches 0.9 times its steady-state value, $K_{res} V_m$, over the time interval known as the *rise time*

$$\tau_r = 2.303\,\tau_{ckt} = 4.606 Q_{eq}/\omega_{res} \tag{9.64}$$

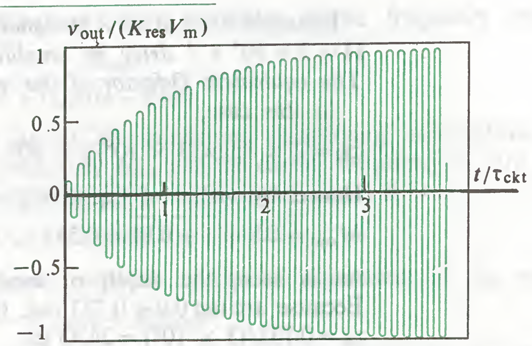


Fig. 9.8 Transient response of a tuned amplifier tuned to the signal frequency

The effect of detuning. Now we will consider the previous problem in a more general form on assuming that the frequency of the harmonic signal differs from the resonant frequency of the tuned circuit by an amount $\delta\omega$:

$$v_{in}(t) = V_m \cos[(\omega_{res} + \delta\omega)t]$$

and

$$\tilde{V}_{in}(t) = V_m \exp(j\delta\omega t)\sigma(t)$$

The simplest way to find the output signal from the l.f. equivalent of a single-stage tuned amplifier is to use the Duhamel superposition integral in which the impulse response of the l.f. equivalent

$$h_{l.f}(t) = -(K_{res}/\tau_{ckt}) \exp(-t/\tau_{ckt})\sigma(t) \tag{9.65}$$

is inserted.

On the basis of Eq. (9.54), we find

$$\tilde{V}_{out}(t) = -\frac{K_{res} V_m}{\tau_{ckt}} \int_0^t \exp(j\delta\omega\tau) \exp[-(t-\tau)/\tau_{ckt}]\, d\tau$$

$$= \frac{-K_{res} V_m}{1 + j\delta\omega\tau_{ckt}} [\exp(j\delta\omega t) - \exp(-t/\tau_{ckt})] \tag{9.66}$$

The physical envelope of the oscillatory waveform appearing at the output of a tuned amplifier is described by the magnitude of the complex output envelope

$$V_{out}(t) = |\tilde{V}_{out}(t)| = \frac{K_{res} V_m}{\sqrt{1 + (\delta\omega\tau_{ckt})^2}}$$

$$\times \sqrt{1 - 2\exp(-t/\tau_{ckt}) \cos \delta\omega t + \exp(-2t/\tau_{ckt})} \tag{9.67}$$

The plots constructed by Eq. (9.67) for several values of the detuning $\delta\omega$ appear in Fig. 9.9.

Thus, the difference between the resonant frequency of the tuned circuit and the harmonic carrier frequency of the output signal results in nonmonotonic variations in the envelope of the output signal.

The physical interpretation of this fact is this: The output signal of an amplifier is the sum of the forced (or steady-state) response which has the frequency of the external source, and the free (or transient) response which exponentially decays with time and has a frequency equal to the natural frequency of the tuned circuit. For the duration of the output signal, the phasor $\dot{V}_{free}$ rotates at the difference frequency relative to the phasor $\dot{V}_{forced}$. The envelope of the output signal, proportional to the length of the resultant phasor

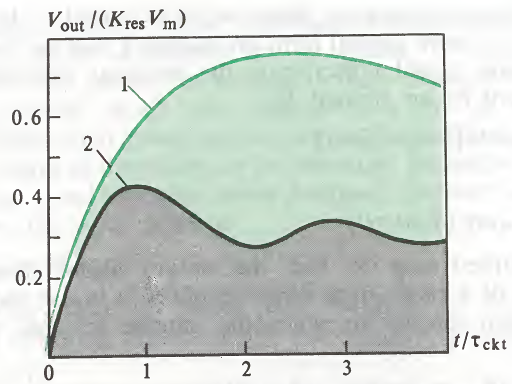
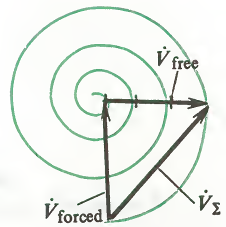


Fig. 9.9 Transients in a tuned amplifier under detuning: (*1*) for $\delta\omega\tau_{ckt}=1$; (*2*) for $\delta\omega\tau_{ckt}=3$

$\dot V_\Sigma$, is time-varying, tending in the limit to the amplitude of the forced (steady-state) response.



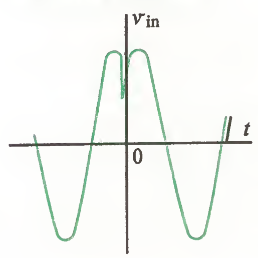
Summation of the free (transient) and forced (steady-state) waves

Interestingly, as the output approaches its steady state, the instantaneous frequency of the output signal is varying, too. Using Eq. (9.66) and taking advantage of the method for finding instantaneous frequency values set forth in Chap. 5, we get

$$\omega(t)=\omega_{res}+\frac{d}{dt}\arg\tilde V_{out}(t)$$

$$=\omega_{res}+\frac{d}{dt}\arctan\frac{\sin\delta\omega t}{\cos\delta\omega t-\exp(-t/\tau_{ckt})} \tag{9.68}$$

Naturally, at $t\to\infty$, when the transients in the amplifie have practically died out, the frequency of the output signal becomes equal to the frequency of the excitation.



PSK signals

Response of a tuned amplifier to a phase-shift-keyed (PSK) drive. As already noted, communication services frequently use truncated harmonic waves whose initial phase undergoes abrupt changes at discrete instants. This form of modulation is known as *phase-shift keying* (PSK), and the corresponding signals are referred to as PSK signals.

Let us take a closer look at the response of a single-stage tuned amplifier to PSK signals, assuming that the excitation experiences a change of phase by φ_0 radians at $t=0$:

$$v_{in}(t)=V_m\begin{cases}\cos\omega_{res}t & \text{for } t<0\\ \cos(\omega_{res}t+\varphi_0) & \text{for } t>0\end{cases} \tag{9.69}$$

The complex envelope corresponding to the above signal is

$$\tilde V_{in}(t)=V_m[\sigma(-t)+\exp(j\varphi_0)\sigma(t)] \tag{9.70}$$

Using the Duhamel superposition integral, we find the complex envelope of the output signal:

$$\tilde V_{out}(t)=\frac{-K_{res}V_m}{\tau_{ckt}}\int_{-\infty}^{t}[\sigma(-\tau)+\exp(j\varphi_0)\sigma(\tau)]$$
$$\times\exp[-(t-\tau)/\tau_{ckt}]\,d\tau \tag{9.71}$$

For $t<0$, it follows from Eq. (9.71) that

$$\tilde V_{out}(t)=\frac{-K_{res}V_m}{\tau_{ckt}}\int_{-\infty}^{t}\exp[-(t-\tau)/\tau_{ckt}]\,d\tau=-K_{res}V_m \tag{9.72}$$

which implies that prior to the jump in phase the amplifier is in a steady state. On the other hand, for $t>0$

$$\tilde V_{out}(t)=-\frac{K_{res}V_m}{\tau_{ckt}}\int_{-\infty}^{0}\exp[-(t-\tau)/\tau_{ckt}]\,d\tau$$
$$-\frac{K_{res}V_m\exp(j\varphi_0)}{\tau_{ckt}}\int_{0}^{t}\exp[-(t-\tau)/\tau_{ckt}]\,d\tau$$
$$=-K_{res}V_m\{\exp(-t/\tau_{ckt})$$
$$+\exp(j\varphi_0)[1-\exp(-t/\tau_{ckt})]\} \tag{9.73}$$

It is to be noted that for $t=0$ Eqs. (9.72) and (9.73) yield the same result:

$$\tilde V_{out}(0)=-K_{res}V_m$$

If, however, $t/\tau_{ckt}\gg1$, then

$$\tilde V_{out}(t)\approx-K_{res}V_m\exp(j\varphi_0)$$

that is, when all transients have died out the system goes into a new steady state which differs from the previous one by a phase shift of φ_0 radians.

From Eq. (9.73) we may write the following expression for the physical envelope of the output signal for $t>0$:

$$V_{out}(t)=K_{res}V_m\{[e^{-t/\tau_{ckt}}+(1-e^{-t/\tau_{ckt}})\cos\varphi_0]^2$$
$$+(1-e^{-t/\tau_{ckt}})^2\sin^2\varphi_0\}^{1/2} \tag{9.74}$$

In practical PSK, use is often made of a phase shift of 180°. Then

$$V_{out}(t)=K_{res}V_m\,|\,2\exp(-t/\tau_{ckt})-1\,| \tag{9.75}$$

PSK with a phase shift of 180° is especially convenient for the transmission of binary-coded messages

Here the amplitude of the output signal goes to zero at time t_0 which is the root of the equation

$$2\exp(-t_0/\tau_{ckt})-1=0$$

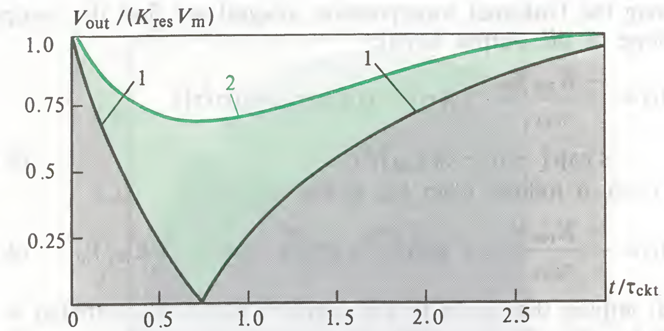


Fig. 9.10 Signal envelope at the output of a tuned amplifier driven by a phase-shift-keyed signal: (*1*) for $\varphi_0 = 180°$; (*2*) for $\varphi_0 = 90°$

Hence,

$$t_0 = 0.693\tau_{ckt} \tag{9.76}$$

The dependence of the physical envelope of the output signal on the dimensionless parameter t/τ_{ckt} for two values of the phase shift φ_0, 180° and 90°, is shown in Fig. 9.10.

The physical behaviour of the output signal is due to the interference of the free and forced oscillations.

The instability of instantaneous frequency is undesirable

It is to be noted that the transients arising in an amplifier in response to a PSK signal are accompanied by variations in the instantaneous frequency. On the basis of Eq. (9.73)

$$\omega(t) = \omega_{res} + \frac{d}{dt}\arctan\frac{(1 - e^{-t/\tau_{ckt}})\sin\varphi_0}{e^{-t/\tau_{ckt}} + (1 - e^{-t/\tau_{ckt}})\cos\varphi_0} \tag{9.77}$$

Response of a resonant system to an angle-modulated signal. An exact solution for the response of a narrowband frequency-selective network to an FM or a PM excitation is difficult to obtain. The l.f. equivalent method set forth in this chapter only permits us to formulate the problem. Thus, if the input to a single-stage tuned amplifier is a simple single-tone angle-modulated wave

$$v_{in}(t) = V_m \cos(\omega_0 t + m \sin \Omega t)$$

whose complex envelope is

$$\dot{V}_{in}(t) = V_m \exp(jm \sin \Omega t)$$

then the complex envelope of the output signal can be described by the following Duhamel superposition integral

$$\dot{V}_{out}(t) = \frac{-K_{res}V_m}{\tau_{ckt}} \int_{-\infty}^{t} \exp(jm\sin\Omega\tau)\exp[-(t-\tau)/\tau_{ckt}]\,d\tau \tag{9.78}$$

(It is assumed that $\omega_0 = \omega_{res}$.) Since the integral in (9.78) has a varying upper limit, it cannot be evaluated exactly. The problem can be solved approximately, if we assume that the instantaneous frequency of the input wave varies so slowly that the oscillatory system can "follow" its variations. Then the envelope of the output signal will be proportional to the magnitude response at the instantaneous frequency:

$$|K(j\omega)| = \frac{K_{res}}{\sqrt{1 + \tau_{ckt}^2(\omega - \omega_{res})^2}} \tag{9.79}$$

On inserting $\omega = \omega_{res} + m\Omega\cos\Omega t$, we obtain the following equation for the envelope of the output signal:

$$V_{out}(t) = \frac{K_{res}V_m}{\sqrt{1 + (m\Omega\tau_{ckt})^2\cos^2\Omega t}} \tag{9.80}$$

Thus, the passage of an FM or a PM signal through a carrier-tuned resonant amplifier is accompanied by a parasitic amplitude modulation. If $m\Omega\tau_{ckt} \ll 1$, then the following approximate expression stems from (9.80):

$$V_{out}(t) \approx K_{res}V_m[1 - \frac{1}{2}(m\Omega\tau_{ckt})^2\cos^2\Omega t]$$

$$= K_{res}V_m\left\{\left[1 + \frac{(m\Omega\tau_{ckt})^2}{4}\right] - \frac{(m\Omega\tau_{ckt})^2}{4}\cos 2\Omega t\right\} \tag{9.81}$$

It means that the envelope spectrum includes the second harmonic of the modulating frequency and, if the frequency deviation is large enough, also the higher even harmonics.

The instantaneous-frequency method

The instantaneous-frequency method permits us not only to analyse variations in the amplitude of FM and PM signals, but also to find the variations caused in the signal by the oscillatory system of the amplifier itself. To this end, we note that the total phase of the output signal is the sum of that of the input signal $\Psi_{in} = \omega_0 t + m\sin\Omega t$

Solve Problem 10

and the phase shift due to the phase response of the amplifier:

$$\varphi_K(\omega) = -\arctan(\omega - \omega_{res})\tau_{ckt} \tag{9.82}$$

On inserting the instantaneous frequency of the input wave

$$\omega_{in} = \omega_{res} + m\Omega\cos\Omega t$$

in Eq. (9.82), we get the total phase Ψ_{out} and the corresponding

instantaneous frequency ω_{out} of the output signal:

$$\Psi_{out} = \omega_0 t + m \sin \Omega t - \arctan(m\Omega\tau_{ckt} \cos \Omega t) \quad (9.83)$$

$$\omega_{out} = d\Psi_{out}/dt = \omega_0 + m\Omega \cos \Omega t + \frac{m\tau_{ckt}\Omega^2 \sin \Omega t}{1 + (m\Omega\tau_{ckt} \cos \Omega t)^2} \quad (9.84)$$

Now we assume that the product of the frequency deviation by the tuned-circuit time constant $b = m\Omega\tau_{ckt} \ll 1$ and expand the last term on the right-hand side of (9.84) into a power series. On retaining the first term of the series, we obtain

$$\omega_{out} = \omega_0 + m\Omega \cos \Omega t + \Omega b \sin \Omega t - \Omega b^3 \sin \Omega t \cos^2 \Omega t$$

By simple trigonometry, we finally get

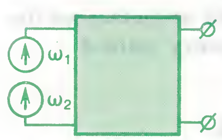
The spectral composition of the instantaneous frequency at the output of a narrowband system

$$\omega_{out} = \omega_0 + m\Omega \cos \Omega t + \Omega b (1 - b^2/4) \sin \Omega t - (\Omega b^3/4) \sin 3\Omega t \quad (9.85)$$

As follows from Eq. (9.85), the instantaneous frequency of the output signal contains not only the in-phase component $m\Omega \cos \Omega t$ which is present in the input signal, but also a small quadrature component which is proportional to $\sin \Omega t$. From its presence we conclude that the signal embedded in the instantaneous frequency has been shifted in time. In fact, the above equation includes one more component which varies at three times the modulating frequency. It is a manifestation of the distorting effect that a narrowband network has on the useful message embedded in the current values of the instantaneous frequency. Fortunately, under the assumptions that we have made this detrimental effect is negligible.

The role of the phase characteristic of a network. Before we conclude our overview of the properties and characteristics of linear stationary networks, it is important to see how the phase response of a system affects its output signal.

For better insight into the fundamental aspects of the phenomena involved, let us consider a case where the excitation is the sum of two harmonic waves. Each has an amplitude of unity, and their frequencies, ω_1 and ω_2, are chosen such that their difference is small:



$$|\omega_1 - \omega_2| / \omega_1 \ll 1$$

Analytically, the input signal may be written as

$$v_{in}(t) = \cos \omega_1 t + \cos \omega_2 t = 2\cos\left(\frac{\omega_1 - \omega_2}{2} t\right) \cos\left(\frac{\omega_1 + \omega_2}{2} t\right) \quad (9.86)$$

Narrowband groups

The energy of the aggregate process is concentrated in time as individual "portions" called *narrowband* or *quasiharmonic groups*. The smaller the difference in frequency between the two input waves, the more these groups are extended in time. The low-frequency term, $2\cos\left(\frac{\omega_1 - \omega_2}{2} t\right)$ in Eq. (9.86) is the group envelope. A narrowband group may be looked upon as one of the simplest elements that make up a wave having a more complicated spectral structure.

Let the signal defined in (9.86) be applied to a linear stationary system whose frequency response is

$$K(j\omega) = |K(j\omega)| \exp[j\varphi_K(\omega)] \quad (9.87)$$

Suppose that within the frequency interval (ω_1, ω_2) the magnitude of the frequency response is a constant, K_0. Then

$$v_{out}(t) \approx K_0 \{\cos[\omega_1 t + \varphi_K(\omega_1)] + \cos[\omega_2 t + \varphi_K(\omega_2)]\} \quad (9.88)$$

The phase response of the system can be expanded into a Taylor series about the point ω_1:

$$\varphi_K(\omega_2) = \varphi_K(\omega_1) + \frac{d\varphi_K}{d\omega}(\omega_2 - \omega_1) + \ldots$$

Retaining only the linear term of the expansion, we get

$$v_{out}(t) \approx K_0 \{\cos[\omega_1 t + \varphi_K(\omega_1)] + \cos[\omega_1 t + \Delta\omega t + \varphi_K(\omega_1) + \frac{d\varphi_K}{d\omega}\Delta\omega]\}$$

$$= 2K_0 \cos\left[\frac{\Delta\omega}{2}\left(t + \frac{d\varphi_K}{d\omega}\right)\right] \cos[\omega_1 t + \frac{1}{2}\Delta\omega t + \varphi_K(\omega_1) + \frac{1}{2}\frac{d\varphi_K}{d\omega}\Delta\omega]$$

Solve Problem 11

Here, $\Delta\omega = \omega_2 - \omega_1$. It immediately follows that the envelope of a quasiharmonic group at the output of the system is shifted in time from the envelope of the input signal by an amount equal to $|d\varphi_K/d\omega|$. If this derivative is negative, the envelope of the output signal is delayed by a time

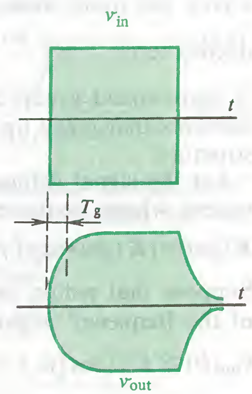
Group (or envelope) delay time

$$T_g = -d\varphi_K/d\omega \quad (9.89)$$

called the *group* (or *envelope*) *delay time.* The derivative must be taken at an arbitrary point within the frequency interval where the bulk of the energy of the narrowband signal is concentrated.

The group (or envelope) delay time is a convenient measure of the delay that narrowband signals experience on passing through sluggish linear circuits.

Example 9.7. *A rectangular r.f. pulse of duration* $\tau_p = 20\ \mu s$ *and with a carrier frequency* $f_0 = 10$ MHz *is applied to a carrier-tuned single-stage resonant amplifier. The equivalent Q-factor of the tuned circuit is* $Q_{eq} = 40$. *Find the delay time of the output pulse relative to the input wave.*



The spectrum of the input signal is concentrated in the frequency interval $(f_0 - 1/\tau_p, f_0 + 1/\tau_p)$, that is, between 9.950 and 10.050 MHz (the spectrum width is taken between the zeros of the main lobe of the spectrum diagram). The bandwidth of the amplifier is

$$BW_{0.707} = f_{res}/Q_{eq} = 250 \text{ kHz}$$

Approximately, it may be taken that the input pulse in our case is a narrowband signal whose delay time may be evaluated by Eq. (9.89). Using the equation of the phase response

$$\varphi_K(\omega) = -\arctan \tau_{ckt}(\omega - \omega_{res})$$

we get

$$T_g|_{\omega=\omega_{res}} = \tau_{ckt} = 2Q_{eq}/\omega_{res} = 1.27\ \mu s$$

Summary

- A salient property of narrowband frequency-selective systems is the fact that for them the ratio of the bandwidth to the centre (midband) frequency is small. These networks are used for the frequency filtering of band-limited signals.
- Physically, frequency-selective networks can be realized as high-Q resonant circuits, both single- and multistage.
- The input signal will be broadband with respect to a frequency-selective system if its spectrum may be considered constant over the pass band.
- The low-frequency equivalent of a narrowband network is an imaginary system whose frequency response is obtained by translating the frequency response of the original network into the neighbourhood of the zero frequency.
- The impulse response of a narrowband network is a narrowband wave whose instantaneous frequency is close to the central frequency of the passband.
- The complex envelope of the impulse response of a narrowband network is proportional to the impulse response of the low-frequency equivalent.
- Frequency-selective networks can be utilized for experimental spectrum analysis in which the response of a frequency-selective network to an arbitrary broadband excitation is measured in order to trace the frequency dependence of the signal spectrum.
- Instead of seeking a complete solution defining the response of a frequency-selective network to a narrowband signal, it will suffice to tackle a simpler problem in which one determines the response of the low-frequency equivalent of the original network to the complex envelope of the input signal.
- If the central frequency of the passband and the carrier frequency are the same, a narrowband network will reduce the modulation depth of an AM input signal.

- When a single-stage tuned system is driven by a truncated sinewave, the envelope of the output process rises to 90% of its final (steady-state) value in a time $\tau_r = 4.606\, Q_{eq}/\omega_{res}$.
- If there is a difference between the resonant frequency of a tuned circuit and the carrier frequency of the driving pulse, the envelope of the output signal will vary nonmonotonically in time. The instantaneous frequency at the output will vary, too.
- The passage of a PSK signal through of a narrowband tuned amplifier produces time variations in both the envelope and the instantaneous frequency of the output signal.
- In an elementary form, the response of narrowband networks to FM and PM inputs can be determined only on the assumption that the product of the input frequency deviation by the system time constant is sufficiently small (the instantaneous-frequency method).
- A frequency-selective network is responsible for spurious amplitude modulation of FM and PM output signals, and also for the appearance of the 3rd harmonic of the modulating frequency in the instantaneous frequency spectrum.

Review Questions

1. What is the convention for defining the bandwidth of narrowband electric networks? What is the attenuation, in decibels, of the signal at the boundary of the bandwidth?
2. What is the condition for the amplitude response of a single-stage resonant system to be symmetrical about the resonant frequency?
3. What is the typical order of magnitude for L, C and R_{res} in parallel resonant circuits with a resonant frequency of several tens of megahertz?
4. Define absolute, relative and generalized detuning.
5. Plot a typical location of the transfer-function poles of a narrowband system in a complex plane. Can the plot be used to find the Q-factor of the system?
6. What is the tapping-down of a tuned circuit? List the salient features of a tapped tuned circuit.
7. How does the internal resistance of an electron device affect the behaviour of a small-signal tuned amplifier? How can one mitigate the resultant detrimental effect? Write equations for finding (1) the gain at resonance and (2) the bandwidth of the amplifier.
8. What is the advantage of a coupled-circuit amplifier over a single-stage tuned amplifier?
9. List the factors which can, in your opinion, determine the upper frequency limit at which a tuned amplifier is still operable.
10. What is the meaning, absolute or relative, of the concept of broadband signal?
11. Is it necessary for the low-frequency equivalent of a narrowband system to be a physically realizable network?
12. Construct an approximate plot for the impulse response of a narrowband system. Note the condition for physical realizability.
13. What is the time constant of a resonant circuit?
14. What is the difference between the impulse responses of a single-stage and a two-stage amplifier?
15. Define the physical meaning of the spectral expansion of a signal.
16. How should the bandwidth of a tuned amplifier be chosen so that AM signals can be

satisfactorily passed in practice? Where do the requirements for the shape of the amplitude response of frequency filter come in conflict?
17. What factors decide the rise time of a single-stage tuned amplifier?
18. Explain in physical terms the rise of the output of a single-stage tuned amplifier to its steady-state value when the amplifier is driven by a truncated voltage sinewave. What is the role played by the free (transient) response of the amplifier? Why is it that initially the output signal is small?
19. Do the same for a PSK input signal. Why is it that the output frequency does not remain constant during transients?
20. Define the rationale of the instantaneous-frequency method.
21. Construct an approximate plot of the wanted signal that can be extracted from the output wave of a single-stage resonant system, assuming that the input signal is a single-tone FM signal.
22. Define the group (envelope) delay time. What should the phase response of a system be like for the applied signal to suffer a minimum of distortion?

Problems

1. Define the value of generalized detuning for which the slope of the amplitude response of a resonant circuit is a maximum.

2. A parallel resonant circuit has a resistance of 30 kΩ at the resonant frequency of 20 MHz, while the magnitude of its impedance at a frequency of 21 MHz is 18 kΩ. Find the element values of the tuned circuit.

3. Find the frequency response and the impulse response of the low-frequency equivalent of an ideal bandpass filter. *Hint*: see Eq. (9.23).

4. Work the previous problem as applied to a Gaussian radio filter. *Hint*: See Eq. (9.24).

5. In the small-signal tuned amplifier of Fig. 9.3, the tapping-down factor for the tuned circuit in the collector lead of the transistor can be varied by shifting the tap along the tuned-circuit coil. Find the relation between the voltage across the tuned-circuit capacitor and the tapping-down factor.

6. A Gaussian radio filter [see Eq. (9.24)], for which $K_0 = 10$, $\omega_0 = 10^6\ \mathrm{s}^{-1}$ and $b = 5 \times 10^{-10}\ \mathrm{s}^2$, is driven by a rectangular video pulse of 25 V amplitude and 0.2 μs duration. Verify that for the filter in question the input signal may be classed as broadband.

7. A single-stage tuned amplifier, for which $f_{res} = 6$ MHz, $Q_{eq} = 40$ and $K_{res} = 35$, is driven by an exponential video pulse of voltage (V)

$$v_{in}(t) = 0.3\exp(-4 \times 10^7 t)\,\sigma(t)$$

Find the output signal of the amplifier.

8. A series resonant circuit is driven by a source of emf (V)

$$v_{in}(t) = 5 \times (1 + 0.8\cos 4 \times 10^3 t)\cos 10^6 t$$

The resonant circuit is tuned to resonate at the carrier frequency. Find the Q-factor of the tuned circuit at which the current modulation factor is 0.4.

9. Analyse distortions in the envelope of rectangular radio pulses on passing through a single-stage tuned amplifier in which the tuned circuit has $Q_{eq} = 60$. The carrier frequency (2 MHz) of the input signal is the same as the resonant frequency of the tuned circuit.

10. Define the law of time variations in the instantaneous frequency of the output signal from a single-stage tuned amplifier driven by a voltage wave

$$v_{in}(t) = V_0\cos(2\pi 10^6 t + 0.1\cos 2\pi 10^3 t)$$

The amplifier tuned circuit is tuned to the carrier frequency of the input signal and has a time constant of 10^{-5} s.

11. A tuned amplifier has two stages tuned to the same frequency of 12 MHz. The Q-factor of the tuned circuit in the 1st stage is 40, and that in the 2nd stage is 50. Find the group (envelope) delay time for a narrowband signal whose centre frequency is 12.2 MHz.

Advanced Problems

12. Find the response of a tuned amplifier to a single-tone AM signal when the carrier frequency ω_0 differs from the resonant frequency ω_{res} of the tuned circuit. Show that the difference between the two frequencies leads to a parasitic angle modulation of the output signal. Assuming that the relative detuning, $|\omega_0 - \omega_{res}|/\omega_0$, is small, derive an equation for the parameters of the parasitic modulation.

13. Investigate the response of a single-stage amplifier to an excitation whose instantaneous frequency undergoes a step change at $t = 0$:

$$v_{in}(t) = \begin{cases} V_0\cos\omega_0 t & \text{for } t < 0 \\ V_0\cos(\omega_0 + \delta\omega)t & \text{for } t > 0 \end{cases}$$

Assume that $\omega_{res} = \omega_0$. Find the expression describing variations in the envelope and instantaneous frequency of the output signal.

14. Suppose a tuned amplifier consists of identical stages. Derive the equation defining the rise time of the system as a function of the number N of stages, stage time constant τ_{ckt}, and resonant frequency ω_{res}. The input signal is a truncated voltage sinewave at frequency $\omega_0 = \omega_{res}$. *Hint:* Take advantage of the fact that

$$\frac{1}{p(p+a)^N} = \frac{1}{a^N}\left[1 - \exp(-at)\sum_{m=0}^{N-1}\frac{(at)^m}{m!}\right]$$

Chapter 10

Response of Linear Stationary Networks to Random Signals

In the previous two chapters, we dealt with methods which permit solving any problems involving the passage of deterministic signals through linear stationary systems. The final step in the theory of linear systems is to extend the above methods to random signals.

Suppose that we have a linear stationary network driven by a wave $x(t)$ which is a realization of a random process $X(t)$. If this realization is specified in advance, no new problem will arise – the signal $x(t)$ may be treated as a sufficiently deterministic function, although it may be described in a fairly complicated manner. Once we know the mathematical model of a system, which may be defined in terms of, say, the frequency response $K(j\omega)$, we can always find the system's response $y(t)$.

A salient feature of statistical communication theory is, however, the fact that so full an information about the input signal is not available – instead of a deterministic description of the input signal we have to be content with the statistical averages of the random process $X(t)$. These statistical averages are the univariate and bivariate probability densities and various moment functions, above all the expectation and the autocorrelation function of the random process. Our objective is to investigate the association between the processes $X(t)$ and $Y(t)$ that can be defined on the basis of the frequency response of the system involved.

10.1 Spectral Analysis of the Response of Linear Stationary Circuits to Random Signals

From the outset it is essential to impose an important constraint – we shall deal solely with wide-sense stationary input processes $X(t)$. As will be recalled, this implies that the expectation (mean) of the instantaneous values of realizations, $\bar{x}$, is time-invariant, whereas the autocorrelation function $K_x(t_1, t_2) = \overline{x(t_1)x(t_2)} - \bar{x}^2$ depends solely on $\tau = |t_1 - t_2|$, the absolute difference between observation times (points on the time axis).

To simplify matters in our subsequent discussion, we shall always put $\bar{x} = 0$. This will not limit the generality of reasoning and conclusions. Owing to the linearity of the circuits in question, the effect of the constant (d.c.) component in the input signal on the response of a system can be analysed independently and, importantly, without recourse to statistical procedures.

The mean of the output signal. Let us take an individual

realization $x(t)$ of the input signal and represent it as a Fourier integral

$$x(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} S_x(\omega)\exp(j\omega t)\,d\omega$$

Here and elsewhere the realization is assumed to be such that its spectrum does exist, at least in the form of a generalized function.

The output signal of the system can be found, if we know its frequency response

$$y(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} S_x(\omega)K(j\omega)\exp(j\omega t)\,d\omega \tag{10.1}$$

Averaging is done over an ensemble of realizations

When changing from an individual realization to a complete ensemble of input signals, we should take $S_x(\omega)$ as a random function and recall that the assumption of stationarity (see Chap. 7) of the process $X(t)$ imposes rigorous conditions: The mean of the spectrum is equal to zero, $\overline{S_x(\omega)} = 0$. Therefore, on averaging statistically both sides of Eq. (10.1), we get

$$\bar{y} = \frac{1}{2\pi}\int_{-\infty}^{\infty} \overline{S_x(\omega)}K(j\omega)\exp(j\omega t)\,d\omega = 0 \tag{10.2}$$

The autocorrelation function and the power spectrum of a random output signal. In order to find the autocorrelation function $K_y(\tau)$, we should, in addition to the spectral expansion (10.1), also have an expression for the output signal at the time $t + \tau$. It can be found on the basis of the properties of the Fourier transform:

$$y(t+\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty} S_x(\omega')K(j\omega')\exp(j\omega'\tau)\exp(j\omega' t)\,d\omega' \tag{10.3}$$

There is a small (and, indeed, an unimportant) point related to the computational technique: the function $y(t)$ is real, and so Eq. (10.3) will remain unaffected if we pass to complex-conjugate quantities on its right-hand side:

$$y(t+\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty} S_x^*(\omega')K^*(j\omega')\exp(-j\omega'\tau)\exp(-j\omega' t)\,d\omega' \tag{10.4}$$

The autocorrelation function of the output signal can now be

found by multiplying together the signals defined by Eqs. (10.1) and (10.4) and averaging the product statistically:

$$K_y(\tau) = \overline{y(t)y(t+\tau)} = \frac{1}{4\pi^2}\int\int_{-\infty}^{\infty} \overline{S_x(\omega)S_x^*(\omega')}K(j\omega)K^*(j\omega')$$

$$\times \exp(-j\omega'\tau)\exp[j(\omega-\omega')t]\,d\omega\,d\omega' \tag{10.5}$$

At first glance the above equation may seem hopelessly complicated to be analysed. But it should be borne in mind that the input random process in question is stationary, and so (see Chap. 7) the random spectra of its individual realizations are delta-correlated, that is,

$$\overline{S_x(\omega)S_x^*(\omega')} = 2\pi W_x(\omega)\,\delta(\omega-\omega') \tag{10.6}$$

The filtering property of the delta function is utilized

where $W_x(\omega)$ is the power spectrum (or the power spectral density) of the stationary random process $X(t)$. From this remarkable feature in the structure of the input signal spectrum, we can readily see the simple meaning of Eq. (10.5):

$$K_y(\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty} W_x(\omega)\,|K(j\omega)|^2\exp(-j\omega\tau)\,d\omega$$

or equally

$$K_y(\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty} W_x(\omega)\,|K(j\omega)|^2\exp(j\omega\tau)\,d\omega \tag{10.7}$$

In effect, Eq. (10.7) contains a complete solution to our problem in terms of correlation theory: *The power spectrum of a random output signal is related to that of the associated input signal as follows:*

Relation between the input and output power spectra of a system

$$W_y(\omega) = W_x(\omega)\,|K(j\omega)|^2 \tag{10.8}$$

In applied problems, one has often to deal with one-sided power spectra, $F_x(f)$ and $F_y(f)$, which are defined solely for positive frequencies f in hertz. Obviously,

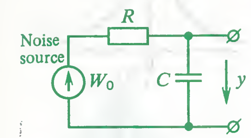
$$F_y(f) = F_x(f)\,|K(j2\pi f)|^2 \tag{10.9}$$

and so the variance of the output signal

$$\sigma_y^2 = K_y(0) = \int_0^{\infty} F_x(f)\,|K(j2\pi f)|^2\,df \tag{10.10}$$

is the result of combining the contributions from the power spectrum of the input signal multiplied by the frequency-dependent square of the magnitude of the frequency response, that is, by the power transfer function.

The manner in which to handle the problems falling in the class examined here is already well known to us – it involves the evaluation of Fourier integrals by a variety of techniques. Therefore in the examples that follow we shall concentrate not so much on the mathematical aspect of the matter, as on the physical features of the processes concerned.



Example 10.1. *The response of an integrating RC-network to white noise.*

Suppose that a first-order dynamic system schematically depicted as an integrating RC-network, is driven by a source of noise emf whose power spectrum W_0 (V^2 s) is constant at all frequencies. Find the variance and the autocorrelation function of the output voltage $y(t)$.

To begin with, we find the power transfer function of the network in question:

$$|K(j\omega)|^2 = \frac{1}{1+\omega^2(RC)^2}$$

Now, on setting $\tau = 0$ and using Eq. (10.7), we find the variance of the output noise:

$$\sigma_y^2 = \frac{W_0}{\pi}\int_0^{\infty}\frac{d\omega}{1+\omega^2(RC)^2} = W_0/2RC \tag{10.11}$$

Work and 2 **Problems 1**

As should be expected, the variance of the output signal decreases with increasing time constant of the network, because this entails a reduction in the band of frequencies effectively passed by the network.

The autocorrelation function of the output signal is

$$K_y(\tau) = \frac{W_0}{2\pi}\int_{-\infty}^{\infty}\frac{\exp(j\omega\tau)\,d\omega}{1+\omega^2(RC)^2} \tag{10.12}$$

Here,

$$\int_{-\infty}^{\infty}\frac{\exp(j\omega\tau)\,d\omega}{1+\omega^2(RC)^2} = 2\int_0^{\infty}\frac{\cos\omega\tau\,d\omega}{1+\omega^2(RC)^2}$$

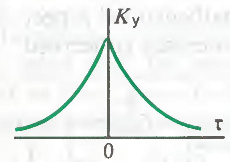
The last integral is tabulated [40], so taking advantage of the fact,

we obtain

$$K_y(\tau) = \frac{W_0}{2RC} \exp\left(\frac{-|\tau|}{RC}\right) \tag{10.13}$$

Naturally, at zero-crossings the autocorrelation function is equal to the variance of the output signal.

To sum up, when an integrating RC-network is excited by white noise, the response is a random process having an exponential autocorrelation function.



Importantly, since it is sluggish in response, the RC-network effects a kind of "ordering": Whereas the input signal is absolutely unpredictable because it is white noise, the output signal is smoothed; its correlation time is of the same order of magnitude as the time constant of the network.

Example 10.2. *The response of a single-stage tuned amplifier to white noise.*

Suppose that a small-signal tuned amplifier is driven by a source of white-noise voltage having the one-sided power spectrum $F_0 (\mathrm{V^2\,Hz^{-1}})$. The frequency response of the system is

$$K(j2\pi f) = \frac{-K_{res}}{1 + j2\pi\tau_{ckt}(f - f_{res})}$$

and the power transfer function is

$$|K(j2\pi f)|^2 = \frac{K_{res}^2}{1 + 4\pi^2\tau_{ckt}^2(f - f_{res})^2}$$

Finding the variance reduces to evaluating the integral in (10.10):

$$\sigma_y^2 = F_0 K_{res}^2 \int_0^{\infty} \frac{df}{1 + 4\pi^2\tau_{ckt}^2(f - f_{res})^2}$$

Now we change the variable, $\eta = f - f_{res}$, and assume that the tuned circuit of the amplifier has so high a Q-factor that the frequency response for $f = 0$ may be taken equal to zero. Then,

$$\sigma_y^2 = F_0 K_{res}^2 \int_{-\infty}^{\infty} \frac{d\eta}{1 + 4\pi^2\tau_{ckt}^2\eta^2} = F_0 K_{res}^2/2\tau_{ckt} \tag{10.14}$$

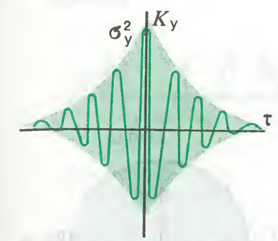
Finally, the autocorrelation function of the output signal is

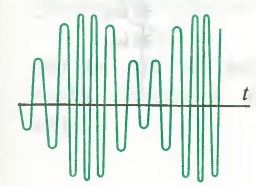
$$K_y(\tau) = \frac{F_0 K_{res}^2}{2\pi} \int_0^{\infty} \frac{\cos \omega\tau \, d\omega}{1 + \tau_{ckt}^2(\omega - \omega_{res})^2}$$

$$\approx \frac{F_0 K_{res}^2}{2\pi} \int_{-\infty}^{\infty} \frac{\cos(\Omega + \omega_{res})\tau \, d\Omega}{1 + \tau_{ckt}^2\Omega^2}$$

$$= \frac{F_0 K_{res}^2}{2\pi} \cos \omega_{res}\tau \int_{-\infty}^{\infty} \frac{\cos \Omega\tau \, d\Omega}{1 + \tau_{ckt}^2\Omega^2}$$

$$= \frac{F_0 K_{res}^2}{2\tau_{ckt}} \exp(-|\tau|/\tau_{ckt}) \cos \omega_{res}\tau \tag{10.15}$$





A typical realization of a random signal at the output of a tuned amplifier

As is seen, the autocorrelation function thus obtained has the form

$$K_y(\tau) = \sigma_y^2 \rho(\tau) \cos \omega_{res}\tau \tag{10.16}$$

which is typical of a narrowband random process, because the envelope $\rho(\tau) = \exp(-|\tau|/\tau_{ckt})$ is a slowly varying function in comparison with the h.f. carrier.

Any realization of the random process at the output of a narrowband amplifier is a quasiharmonic wave with a random envelope and a random instantaneous frequency; on the average, the carrier frequency is equal to the resonant frequency of the tuned circuit. This behaviour of the output signal can readily be understood if we note that the signal results from combining a huge number of elementary responses each of which is proportional to the impulse response of the system (the Duhamel superposition integral principle).

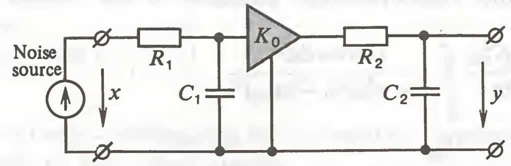
▲ **Solve Problem 4**

In order to get an idea about the order of magnitudes involved in statistical communication, let us estimate the variance of noise at the output of a tuned amplifier under the following conditions: $F_0 = 10^{-16}\ \mathrm{V^2\,Hz^{-1}}$, $K_{res} = 30$, $\omega_{res} = 10^8\ \mathrm{s^{-1}}$, and $Q_{eq} = 60$. The time constant of the tuned circuit is $\tau_{ckt} = 2Q_{eq}/\omega_{res} = 1.2\ \mu\mathrm{s}$. Therefore, on the basis of Eq. (10.14),

$$\sigma_y^2 = 10^{-16} \times 900/(1.2 \times 10^{-6}) = 7.5 \times 10^{-8}\ \mathrm{V^2}$$

The rms value of noise voltage, equal to the square root of the variance, is 274 μV.

Example 10.3. *Consider a network composed of two RC-sections between which an ideal amplifier of gain K_0 is interposed:*



Let the system be excited by a source of noise emf having a power spectrum constant at all frequencies (white noise). Find the autocorrelation function of the output signal.

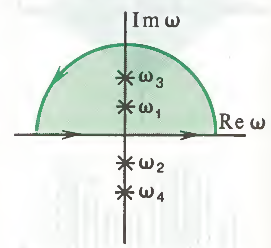
The power transfer function is

$$|K(j\omega)|^2 = \frac{K_0^2}{(1+\omega^2\tau_1^2)(1+\omega^2\tau_2^2)}$$

In accordance with Eq. (10.7), finding the autocorrelation function for the output voltage reduces to evaluating the integral:

$$K_y(\tau) = \frac{W_0 K_0^2}{2\pi} \int_{-\infty}^{\infty} \frac{\exp(j\omega\tau)\,d\omega}{(1+\omega^2\tau_1^2)(1+\omega^2\tau_2^2)} \qquad (10.17)$$

It is advisable to resort to the theory of residues and to take the same route as was used in Chap. 8 in analysing the impulse response of an RC-network. The integrand function in (10.17) has four simple poles at points $\omega_{1,2} = \pm j(1/\tau_1)$ and $\omega_{3,4} = \pm j(1/\tau_2)$.



We will evaluate the function $K_y(\tau)$ for $\tau > 0$ by closing the contour of integration in the upper half-plane. The residue of the integrand function at point ω_1 is

$$\operatorname{res}|_{\omega=\omega_1} = \left.\frac{\exp(j\omega\tau)}{\dfrac{d}{d\omega}[(1+\omega^2\tau_1^2)(1+\omega^2\tau_2^2)]}\right|_{\omega=\omega_1} = \frac{\tau_1 \exp(-\tau/\tau_1)}{2j(\tau_1^2-\tau_2^2)}$$

In a similar way, we find the residue for point $\omega = \omega_3$ lying within the contour of integration

$$\operatorname{res}|_{\omega=\omega_3} = \frac{-\tau_2 \exp(-\tau/\tau_2)}{2j(\tau_1^2-\tau_2^2)}$$

Hence, by virtue of Cauchy's residue theorem, we obtain

$$K_y(\tau) = \frac{W_0 K_0^2}{2(\tau_1^2-\tau_2^2)}[\tau_1 \exp(-\tau/\tau_1) - \tau_2 \exp(-\tau/\tau_2)] \qquad (10.18)$$

for $\tau > 0$.

We can obtain the autocorrelation function for $\tau < 0$ from the same equation on replacing τ with $-\tau$. This stems from the fact that the autocorrelation function is an even function. However, the result can be confirmed by direct calculation, if we close the path of

integration with an arc of an infinite radius in the lower half of the complex ω-plane.

Thus,

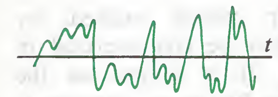
Solve Problem 6

$$K_y(\tau) = \frac{W_0 K_0^2}{2(\tau_1^2-\tau_2^2)}[\tau_1 \exp(-|\tau|/\tau_1) - \tau_2 \exp(-|\tau|/\tau_2)] \qquad (10.19)$$

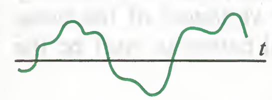
The variance of the output signal is

$$\sigma_y^2 = K_y(0) = \frac{W_0 K_0^2}{2(\tau_1+\tau_2)} \qquad (10.20)$$

The white noise transformed by one RC-network



and by two RC-networks connected in series



Therefore, the correlation coefficient is

$$R_y(\tau) = \frac{1}{\tau_1-\tau_2}[\tau_1 \exp(-|\tau|/\tau_1) - \tau_2 \exp(-|\tau|/\tau_2)] \qquad (10.21)$$

The results obtained by Eq. (10.21) for two values of the ratio τ_1/τ_2 are plotted in Fig. 10.1. For comparison, the same figure shows a curve which represents the correlation coefficient of a random process at the output of a single integrating RC-network with time constant τ_1. It is interesting to note not only the increase in the correlation time due to the addition of a second time-lag element, but also the changed behaviour of $R_y(\tau)$ for the two-section filter in the vicinity of point $\tau = 0$. The fact that the autocorrelation function has a second derivative at that point assures the differentiability of the output random process (see Chap. 7). Physically, the differentiability of the output signal implies the smoothness of the realizations of the signal that has passed through two RC-networks connected in cascade.

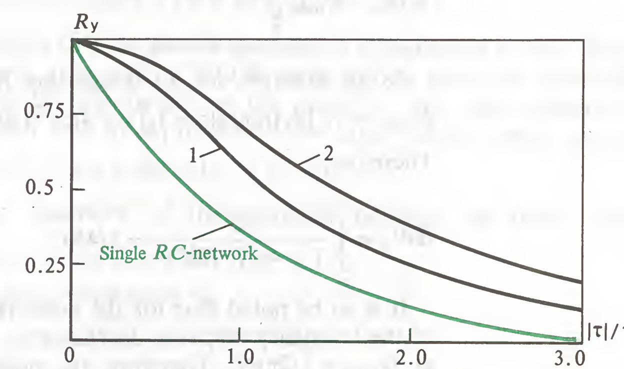


Fig. 10.1 Correlation coefficient for a random process at the output of a system consisting of two RC networks: (1) for $\tau_2 = \tau_1/2$ and (2) for $\tau_1 = \tau_2$

Response of a network to broadband random signals. The noise bandwidth of a network. Frequently one is concerned with the response of linear circuits to very broadband random signals formed by, say, a chaotic series of short pulses. In such cases, there is every ground for considering the spectral characteristics of the noise only within the bandwith of the system, with the real random process replaced by an equivalent white noise having a one-sided power spectrum $F_0 = F_x(f_0)$, where f_0 is a point within the bandwidth of the network.

In the circumstances, Eq. (10.10) which defines the variance of the response can be simplified as

$$\sigma_y^2 = F_0 \int_0^\infty |K(j2\pi f)|^2 \, df \tag{10.22}$$

Noise bandwidth

In engineering calculations, a linear circuit excited by a broadband random signal can conveniently be characterized in terms of its *noise bandwidth*, BW_n (Hz). It is defined as the bandwidth of an ideal bandpass filter whose frequency response is K_{max}, that is, equal to the maximum magnitude of the frequency response of the real network. When an ideal system is excited by white noise with a power spectrum F_0, the variances of the noise signals at the output of the ideal and the real networks must be the same

$$F_0 \int_0^\infty |K(j2\pi f)|^2 \, df = F_0 K_{max}^2 BW_n \tag{10.23}$$

Hence,

$$BW_n = \frac{1}{K_{max}^2} \int_0^\infty |K(j2\pi f)|^2 \, df \tag{10.24}$$

As an example, for an integrating RC-network:

$K_{max} = 1$; $|K(j2\pi f)|^2 = 1/[1 + 4\pi^2 f^2 (RC)^2]$

Therefore,

$$BW_n = \int_0^\infty \frac{df}{1 + 4\pi^2 f^2 (RC)^2} = 1/4RC \tag{10.25}$$

It is to be noted that for the network in question the magnitude of the frequency response decreases to $1/\sqrt{2}$ of its maximum value at $f_{0.707} = 1/2\pi RC$. Therefore, the noise bandwidth is wider than $BW_{0.707}$:

$BW_n/BW_{0.707} = \pi/2 = 1.571$

The noise bandwidth of a single-stage tuned amplifier is found in a similar way:

$$BW_n = \int_0^\infty \frac{df}{1 + 4\pi^2 \tau_{ckt}^2 (f - f_{res})^2} = 1/2\tau_{ckt} = 1.571\, BW_{0.707} \tag{10.26}$$

The example that follows will demonstrate how the concept of noise bandwidth can be used in engineering calculations.

Example 10.4. *A single-stage tuned small-signal amplifier for which $K_{res} = 85$, $Q_{eq} = 70$, and $f_{res} = 0.7$ MHz, is excited by a source of stationary noise emf $x(t)$ for which the autocorrelation function* (V^2) *is*

$K_x(\tau) = 0.45 \exp(-\beta|\tau|)$

where $\beta = 10^7\ s^{-1}$. Find the rms noise voltage at the amplifier output.

To begin with, we invoke the W-K theorem in order to find the power spectrum of the input signal:

$$W_x(\omega) = \int_{-\infty}^{\infty} K_x(\tau) \exp(-j\omega\tau)\, d\tau = 2\sigma_x^2 \int_0^\infty \exp(-\beta\tau) \cos \omega\tau \, d\tau$$

This type of power spectrum has already been found in Chap. 7. Taking advantage of that result, we may write for the case on hand:

$$W_x(\omega) = \frac{2\sigma_x^2 \beta}{\beta^2 + \omega^2} = 0.9 \times 10^7/(10^{14} + \omega^2)$$

The one-sided power spectrum is

$F_x(f) = 2W_x(2\pi f) = 1.8 \times 10^7/[10^{14} + (2\pi f)^2]$

Noting that the power spectrum is a maximum at zero frequency, whereas at $\omega = 10^7\ s^{-1}$ it is only half as great, we conclude that within the bandwidth of the amplifier the input signal may be replaced with an equivalent white noise whose power spectrum is

$F_0 = F_x(f_{res}) = 1.508 \times 10^{-7}\ V^2\ Hz^{-1}$

Solve Problem 8

The bandwidth of the amplifier between the 0.707 points is

$BW_{0.707} = f_{res}/Q_{eq} = 10$ kHz

The noise bandwidth is

$BW_n = (\pi/2) BW_{0.707} = 15.708$ kHz

On substituting the numerical values thus found in Eq. (10.23), we obtain that

$\sigma_y^2 = 1.508 \times 10^{-7} \times 85^2 \times 15708 = 17.11\ V^2$

Hence, the rms value of noise voltage at the amplifier output is

$\sigma_y = \sqrt{17.11} = 4.14$ V

Normalization of a random signal at the output of a linear stationary network. So far we solved all problems within the framework of correlation theory, that is, by invoking at most second-order moment functions. More completely the problem could be stated like this: The input random process is specified by giving a family of its nth-order multivariate probability densities $p_n(x_1, \ldots, x_n; t_1, \ldots, t_n)$. Given the frequency response $K(j\omega)$, it is required to determine the corresponding probability densities for the output process.

▲ Solve Problem 5

In the general case, the solution of such a problem is a highly complicated matter and is not considered in this book. Fortunately, however, we may often expect that the output signal will be Gaussian, irrespective of the probability density of the input process.

The normalization of the output signal is inherent in any stationary linear system if it displays a sufficient time lag (or memory). The point is that in accord with the Duhamel superposition integral

$$y(t) = \int_{-\infty}^{t} x(\tau)\, h(t-\tau)\, d\tau$$

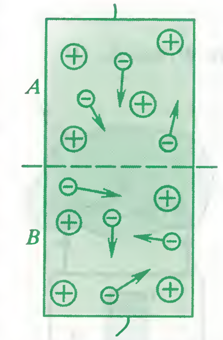
the instantaneous value $y(t)$ is the result of the weighted summation of the previous values of the input signal $x(t)$, multiplied by the time-shifted impulse response (the weighting function) of the network. If the duration of the impulse response is such that it spans several correlation times of the input random process, the conditions for the applicability of the Central Limit Theorem (see Chap. 6) are realized. As a consequence, the output signal shows an asymptotic normality. If, on the other hand, the input random process is normal, the random process at the output of the system will be normal, too, irrespective of the dynamic properties of the system.

Thus, the techniques of correlation theory are quite adequate for solving most problems involving the response of linear stationary circuits to random signals.

10.2 Sources of Fluctuation Noise in Circuit Components

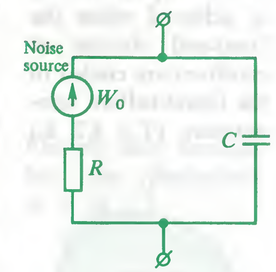
This closing section of Chapter 10 will be concerned with the physical phenomena that stand behind fluctuation voltages and currents in circuits. We will also derive the relations that are used to assess the intensity of noise.

Thermal noise in resistors. The most common cause of noise is the fluctuation of the bulk density of electric charge in conducting bodies (resistors) in turn brought about by the chaotic thermal motion of charge carriers. Although the system as a whole is electrically neutral, time-varying electromagnetic fields are produced in the bulk of the resistor, and a noise potential difference is produced across the external terminals. The noise voltage has an extremely wide spectrum because the charge carriers are very densely packed and the mean thermal velocity is high. In view of this, it is legitimate to think that at radio frequencies the thermal noise voltage across a resistor is an approximate realization of white (delta-correlated) noise.



The charge in region A is not equal to that in region B due to the chaotic motion of the carriers

The Nyquist law. Let us derive the relation for the power spectrum of the noise voltage between the terminals of a resistor R which is in thermal equilibrium with the surroundings at absolute temperature T. To this end, we mentally connect the resistor in parallel with a capacitor C and replace the real noisy resistor with a series combination of an ideal noise-free resistor R and an e.m.f. source producing white noise. It is required to determine the power spectrum W_0 of this noise.



Statistical mechanics tells us that when a system is in thermal equilibrium, its thermal energy is so distributed among the degrees of freedom that each takes on an average amount equal to $kT/2$ (where $k = 1.38 \times 10^{-23}$ J K^{-1} is Boltzmann's constant). This statement is known as the *Law of Equipartition of Energy*. Since the network we are considering is a first-order dynamic system with one degree of freedom, the average electric-field energy stored in the capacitor is exactly $kT/2$, that is,

$$C\overline{v_C^2}/2 = kT/2$$

and so the variance of the noise voltage across the capacitor is

$$\sigma^2 = \overline{v_C^2} = kT/C$$

▲ Solve Problem 7

Now we will use Eq. (10.11) which relates the variance to the power spectrum of the noise voltage. Since

$$W_0/2RC = kT/C$$

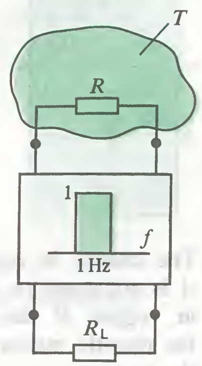
the auxiliary quantity C cancels out, and we get

$$W_0 = 2kTR$$

Practically, it is more convenient to use the one-sided power spectrum defined for the positive frequencies and having the units of V^2 Hz^{-1}:

$$F_0 = 4kTR \tag{10.27}$$

▲ Solve Problem 10



The best performance is achieved when the front-end circuits of receivers are cooled to the liquid-helium temperature ($T = 4.2$ K)

This remarkable relation, known as the *Nyquist Law*, was proved in the late 20s.

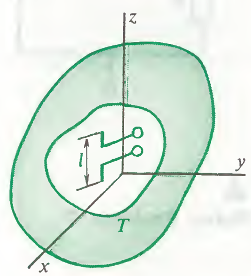
It is to be stressed that F_0 has the simple and clear physical significance of the *specific variance* of the thermal noise source, that is, the variance per frequency band of 1 Hz.

The order of magnitude for the power spectral density of thermal noise can be assessed from the following example. For $T = 300$ K and $R = 10$ kΩ, F_0 will be 1.66×10^{-16} V^2 Hz^{-1}. Hence, the specific rms noise voltage is 1.29×10^{-8} V Hz$^{-1/2}$. Although it appears very small, thermal noise may well be a decisive factor in limiting the actual sensitivity of receivers.

It is interesting and important to note that the noise power that can be transmitted into an external resistive load is independent of R. To demonstrate, consider a system in which an ideal filter with a passband of 1 Hz is placed between R and R_L. As will be recalled from circuit theory, the power transferred into load is a maximum when $R = R_L$ (the matching condition) and is equal in our case to

$$P_{sp} = \overline{v_{sp}^2}/4R = kT \text{ (W Hz}^{-1}) \tag{10.28}$$

Accordingly, the only effective way to minimize thermal noise is to cool the input circuits of sensitive receivers used in radar, radio astronomy and deep-space communications to very low temperatures.



Noise in receiving antennas. In communication equipment an important and, sometimes, the decisive source of noise may be chaotic fluctuations of electromagnetic fields generating noise voltage at the output terminal of a receiving antenna.

Let an elementary receiving antenna (a Hertzian dipole) of length l be oriented along the z-axis and placed inside a closed space whose walls are at a temperature T.

● Specific brightness

By the Planck law, the closed space is filled with an equilibrium electromagnetic radiation characterized by a special spectral parameter known as *specific brightness* and having the units of W m^{-2} Hz^{-1} srad^{-1}:

$$B = \frac{2hf^3}{c^2[\exp(hf/kT) - 1]} \tag{10.29}$$

where c is the velocity of light, f is the frequency in Hz, and $h = 6.62 \times 10^{-34}$ J Hz^{-1} is Planck's constant. Specific brightness is a flux of electromagnetic radiation per Hz, incident at a given point from within a solid angle of 1 steradian.

If $hf \ll kT$ (which is typical of radio frequencies), then Eq. (10.29) transforms into an approximate relation known as the *Rayleigh-Jeans radiation formula*

$$B = 2kT/\lambda^2 \tag{10.30}$$

where $\lambda = c/f$ is the wavelength.

Let us include the quantity

$$\overline{E_{sp}^2} = \overline{E_{x,\,sp}^2} + \overline{E_{y,\,sp}^2} + \overline{E_{z,\,sp}^2}$$

which is the mean square of the electric field strength E per Hz. It is proved in the theory of electromagnetism that in the circumstances the radiation power flux (W m^{-2}) is

$$\overline{E_{sp}^2}/120\pi = \overline{E_{z,\,sp}^2}/40\pi$$

The above expression takes into account the fact that because all the spatial directions are equally significant, $\overline{E_{sp}^2} = 3\overline{E_{z,\,sp}^2}$. By dividing the power flux into 4π, that is, by the solid angle of the entire space, we obtain an expression for brightness in terms of field quantities:

$$B = \overline{E_{z,\,sp}^2}/160\pi^2 \tag{10.31}$$

On equating the right-hand sides of Eqs. (10.30) and (10.31), we find the specific mean square of the electric-field component which is oriented along the antenna:

$$\overline{E_{z,\,sp}^2} = 320\pi^2 kT/\lambda^2 \tag{10.32}$$

Electric field strength has the dimensions of V m^{-1}

Since the voltage induced across the terminals of the antenna (which is short in comparison with the wavelength) is $v = El$, the specific variance of the output voltage is found to be

$$\overline{v_{sp}^2} = 320\pi^2 (l/\lambda)^2 kT \tag{10.33}$$

If we introduce the so-called *radiation resistance* (Ω) of a Hertzian dipole

$$R_\Sigma = 80\pi^2 (l/\lambda)^2$$

then Eq. (10.33) transforms into the Nyquist equation for an elementary receiving antenna:

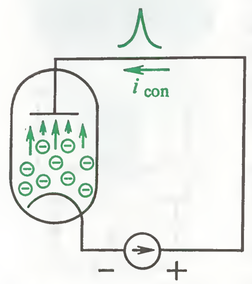
$$\overline{v_{sp}^2} = F_0 = 4kTR_\Sigma \tag{10.34}$$

▲ Solve Problem 11

Here T is the temperature of the equilibrium medium through which electromagnetic waves are propagated. This fully holds only for noise of cosmic origin. As measurements have shown, the "coldest" regions of the sky have a temperature of several kelvins. On the other hand, the temperature in the direction of radio galaxies and other natural sources of radio noise emission may be as high as 10 000 K. As far as atmospheric interference of terrestrial origin is concerned, here the bulk of noise power is concentrated at frequencies below 30 MHz. In order to retain Eq. (10.34) unchanged, one has to introduce the noise temperature T_n which is frequency-dependent. The spectral composition of atmospheric interference is such that at frequencies of the order of 1 MHz the noise temperature may sometimes be as high as 3×10^8 K.

Shot noise derives its name from the resemblance that it bears to that of fine shot falling on a surface.

Shot noise. In this case, very common in electronic circuits, noise is generated due to the passage of discrete charge carriers. In contrast to thermal noise, the fluctuations in this case arise not from the chaotic thermal motion of electrons, but owing to the statistical independence of carrier motions in electron devices (tubes, crystal diodes, transistors, etc.).

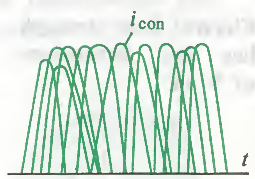


As an example, consider a thermionic (hot-cathode) diode in which a current is caused to flow by a source of constant e.m.f. This current takes the form of a chaotic stream of electrons each of which carries a charge $e = 1.6 \times 10^{-19}$ C. The time for the transit of an electron from the cathode to the anode is $\tau_t \approx 10^{-9}$ s. During this interval, a short pulse of the so-called *convection current* is registered in the external circuits of the diode. The charge thus transferred by each electron is the electronic charge e and so:

$$\int_0^{\tau_t} i_{con}\,dt = e$$

Hence, we estimate that $i_{con} \approx 1.6 \times 10^{-10}$ A.

Ordinarily, the diode current is a few milliamperes, so the pulses of convection current densely overlap in time.



Pulse overlap

Electrons are emitted by the cathode at statistically independent instants. Hence it is clear that the instantaneous value of anode current does not remain constant—it experiences certain fluctuations. In this sense, an electron device acts as a source of a special kind of noise which has come to be known as *shot noise*.

Poisson distribution. Let ν designate the average number of electrons arriving at the anode every second. Experiments have convincingly proved that this numeric characteristic is statistically constant, that is, stationary. The d.c. component of anode current is connected to the parameter ν by a simple relation

$$I_0 = e\nu$$

The number ν is very large: for $I_0 = 1$ mA it is estimated to be $\approx 10^{16}$ s^{-1}.

Before we analyse the process statistically, we will make a simplifying assumption, of minor importance physically, but facilitating calculations: Suppose that in their travel from cathode to anode electrons follow one another singly, so that the probability of electrons arriving at the anode in twos, threes, and so on, is negligible.

The above reasoning also applies to the injection of carriers in semiconductor devices

From the discrete mechanism of the process, we may legitimately consider that if A is the arrival of an electron at the anode in the interval $(t, t+\Delta t)$, then, to within small terms of the order of $(\Delta t)^2$, the probability of this event is given by

$$P_A = \nu\Delta t \qquad (10.35)$$

A complex (or compound) event is formed by two or more independent events

Let $P_0(t)$ designate the probability that no one electron will arrive at the anode during the time interval from zero to t. Then $P_0(t+\Delta t)$ will be the probability of a complex event: Not a single electron will arrive at the anode in either the interval $(0, t)$ or in the interval $(t, t+\Delta t)$. According to the properties of the probability of a compound event, we have

$$P_0(t+\Delta t) = P_0(t)(1 - \nu\Delta t)$$

By passing to the limit for $\Delta t \to 0$, we obtain the differential equation

$$dP_0/dt = -\nu P_0$$

subject to the obvious initial condition $P_0(0) = 1$. The solution of the equation is elementary:

$$P_0(t) = \exp(-\nu t)$$

Since $\nu \approx 10^{16}$ s^{-1}, the probability that not a single electron will arrive at the anode during the time interval of 1 s long will be $\exp(-10^{16})$, which may legitimately be taken as the probability of an improbable event.

Let us analyse the probability $P_1(t)$ that exactly one electron arrives at the anode. In the time interval $(0, t+\Delta t)$ this probability is the sum of the probabilities of two disjoint events:

(a) an electron arrives in the interval $(0, t)$;

(b) an electron arrives in the interval $(t, t+\Delta t)$.

By the rule for the addition of probabilities,

$$P_1(t+\Delta t) = P_1(t)(1 - \nu\Delta t) + P_0(t)\nu\Delta t$$

Hence, we obtain the following differential equation:

$$dP_1/dt = -\nu P_1 + \nu P_0$$

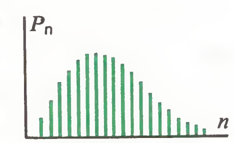
subject to the initial condition $P_1(0) = 0$.

In a similar way, we obtain a solution for the initial-value problem involving the arrival of exactly n electrons at the anode:

$$\begin{cases} dP_n/dt = -\nu P_n + \nu P_{n-1} \\ P_n(0) = 0 \end{cases} \qquad (10.36)$$

By direct substitution, we find that

$$P_n(t) = \frac{(\nu t)^n}{n!}\exp(-\nu t) \qquad (10.37)$$



The Poisson distribution

Equation (10.37) defines the *Poisson distribution* which is encountered in many problems of statistical communication theory. If an experiment yields a Poisson-distributed whole-numbered

variable, variations from the mean in either direction will occur rarely on any one trial.

Moments of a Poisson-distributed random variable. Let us choose a time interval T and count the average number of electrons arriving at the anode:

$$\bar{n}_T = \sum_{n=0}^{\infty} n\frac{(\nu T)^n}{n!}\exp(-\nu T) = \nu T\exp(-\nu T)\sum_{n=1}^{\infty}\frac{(\nu T)^{n-1}}{(n-1)!} = \nu T \tag{10.38}$$

The meaning of the above equation is simple to see: It confirms the original supposition that ν is the mean arrival rate of electrons. Now let us find the mean square of the number of electrons arriving during the chosen time interval:

$$\overline{n_T^2} = \sum_{n=0}^{\infty} n^2\frac{(\nu T)^n}{n!}\exp(-\nu T)$$

$$= \sum_{n=0}^{\infty}[n(n-1)+n]\frac{(\nu T)^n}{n!}\exp(-\nu T)$$

$$= \nu T + (\nu T)^2\sum_{n=2}^{\infty}\frac{(\nu T)^{n-2}}{(n-2)!}\exp(-\nu T) = \nu T + (\nu T)^2 \tag{10.39}$$

Using Eqs. (10.38) and (10.39), we obtain the variance of the number of arriving electrons

$$\sigma_n^2 = \overline{n_T^2} - (\bar{n}_T)^2 = \nu T \tag{10.40}$$

Statistical properties of the diode current. Now let us determine the current flowing through a thermionic diode. If a number n of electrons arrives at the anode over a time T, then the current i_T over the observation interval will be defined by

$$i_T = en/T$$

The mean value of the observed current is

$$I_0 = \bar{i}_T = (e/T)\bar{n}_T = e\nu \tag{10.41}$$

and the variance of the current is

$$\sigma_i^2 = (e^2/T^2)\sigma_n^2 = (e/T)I_0 \tag{10.42}$$

If we choose to measure the intensity of current fluctuations in

terms of the ratio of the rms value to the mean value, then

$$\sigma_i/I_0 = \sqrt{e/T}/\sqrt{I_0} \tag{10.43}$$

The conclusion stemming from Eq. (10.43) is this: The relative fluctuations of the diode current fall off with an increase in both the observation time and the mean current.

Example 10.5. *Let $I_0 = 10^{-2}$ A and the observation time $T = 1$ s.* Then the variance of the current over the observation time is

$$\sigma_i^2 = 1.6\times10^{-21}\ \text{A}^2$$

and the current through the diode will be

$$i = 10^{-2} \pm 4\times10^{-11}\ \text{A}$$

It is seen that the relative fluctuations of current are very small. If we substantially reduce both the mean current and the observation time by putting $I_0 = 10^{-8}$ A and $T = 10^{-8}$ s, then

$$\sigma_i^2 = 1.6\times10^{-19}\ \text{A}^2$$

Work Problem 12

and

$$i = 10^{-8} \pm 4\times10^{-10}\ \text{A}$$

that is, the relative fluctuations of current build up markedly.

The above example explains why we do not notice the random oscillations of the pointer of a moving-coil instrument connected in series with a thermionic diode. Here, the current being measured is averaged over a time interval of about one second due to the sluggishness of the mechanical system of the instrument.

Schottky's equation. The shorter the observation time T, the broader the bandwidth that has to be considered in the spectrum of the process. By Kotelnikov's theorem, for a system to be able to process a signal over a time T, the system must pass all frequencies up to the frequency f_{max} such that

$$T = 1/2f_{max}$$

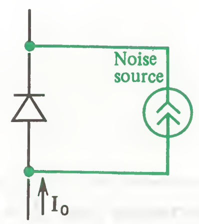
Taking advantage of this in Eq. (10.42), we get

$$\sigma_i^2 = 2eI_0f_{max}$$

Hence, the specific variance of the fluctuating current, that is, the variance per hertz of the frequency band, $\text{A}^2\,\text{Hz}^{-1}$, is

$$\overline{i_{sp}^2} = 2eI_0 \tag{10.44}$$

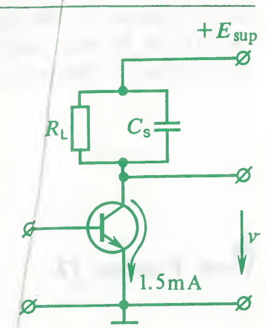
In radio engineering, this important relation has come to be known as *Schottky's equation*. According to it, the equivalent noise circuit



of any electron device contains a current source producing white noise with a power spectrum defined by Eq. (10.44).

Experiments show that the power spectrum of shot noise in electron devices remains flat (that is constant) up to frequencies of several hundred megahertz and then begins to decrease with rising frequency. This is because at very high frequencies (or, which is the same, at small values of T), the adopted noise model, which requires a sufficiently large number of electrons to arrive at the anode over the observation period, is no longer valid. Also, the decrease in the power spectrum of noise due to the finite duration of the elementary convection current pulse begins to take effect.

Example 10.6. *A single-stage RC-coupled transistor amplifier has the following parameters:* $R_L = 5.1$ kΩ, $C_b = 45$ pF, *and* $R_i = 20$ kΩ. *The Q-point is located on the transistor characteristic so that the direct component of the collector current is* $I_0 = 1.5$ mA. *Find the rms output noise voltage due to the shot effect in the transistor.*

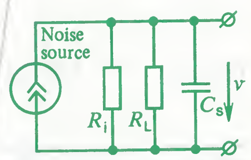


To begin with, we use the Schottky equation to find the power spectrum of the noise current source:

$$F_0 = 2eI_{0,c} = 2 \times 1.6 \times 10^{-19} \times 1.5 \times 10^{-3}$$

$$= 4.8 \times 10^{-22}\ \mathrm{A^2\,Hz^{-1}}$$

In setting up an equivalent output circuit for the amplifier, we note that here the system's frequency response is the complex impedance placed in parallel with the current source:



$$Z(j\omega) = \frac{R_{eq}}{1 + j\omega C_b R_{eq}}$$

where $R_{eq} = R_L R_i/(R_L + R_i)$. In our case,

$$R_{eq} = 5.1 \times 20/25.1 = 4.06\ \mathrm{k\Omega}$$

The variance of the output noise voltage can be found by Eq. (10.22)

$$\sigma_v^2 = F_0 \int_0^\infty |Z(j2\pi f)|^2\, df = F_0 R_{eq}^2 \int_0^\infty \frac{df}{1 + 4\pi^2 R_{eq} C_b f^2}$$

$$= \frac{F_0 R_{eq}}{4C_b} = 1.16 \times 10^{-8}\ \mathrm{V^2}$$

Finally, the rms value of the output noise voltage is

$$\sigma_v = \sqrt{\sigma_v^2} = 108\ \mu\mathrm{V}$$

Such calculations have to be performed whenever one is to determine the sensitivity limit of small-signal amplifiers. The usual procedure is to refer the rms noise voltage to the input by the

equation

$$\sigma_{in} = \sigma_{out}/K$$

where K is the gain of the system at the signal frequency. The value thus obtained is taken to be the least rms voltage of the useful harmonic signal that can be amplified by the device in question:

$$V_{m,s}^{min} = \sqrt{2}\,\sigma_{in}$$

In our example where the transconductance of the transistor is $g_m = 20$ mA/V, the gain at the zero frequency is

$$K = g_m R_{eq} = 81.2$$

Therefore,

$$\sigma_{in} = 108/81.2 = 1.33\ \mu\mathrm{V}$$

and the minimum amplitude of the signal that can be amplified is

$$V_{m,s}^{min} = 1.33 \times \sqrt{2} = 1.88\ \mu\mathrm{V}$$

(It is assumed that the signal frequency is substantially lower than the upper frequency limit of the amplifier.)

Summary

- If the mean value of a stationary random signal at the input of a linear system is zero, then the mean value of the output signal will also be zero.
- The power spectrum of the output random signal is equal to that of the input signal multiplied by the square of the magnitude of the frequency response.
- An integrating RC-network excited by white noise has an exponential autocorrelation function.
- The output signal of a tuned amplifier driven by white noise is a narrowband random process.
- If a linear stationary circuit is sufficiently sluggish, then its output random process is asymptotically normal, irrespective of the statistical properties of the input random process.
- In resistors, thermal noise arises owing to random fluctuations in the electric charge density. The thermal noise of a resistor has a practically constant power spectrum at all frequencies in the r.f. range.
- The power spectrum of thermal noise is found by the Nyquist law.
- The equivalent noise temperature of receiving antennas depends on the frequency band involved and ranges from several kelvins to millions of kelvins.
- The shot effect in electron devices is related to the discrete nature of electron emission and statistical independence between the motions of individual charge carriers.
- The statistical properties of an electron stream are described by the Poisson distribution.
- In an equivalent circuit, an electron device displaying the shot effect may be replaced with a noise current source having a uniform spectrum; the power spectrum of the current is defined by Schottky's equation.

Review Questions

1. How does the stationarity of the input random process figure in the derivation of the equation defining the power spectrum of the output random process?
2. What is the approximate shape of plots for the autocorrelation functions of random signals: (a) at the output of an integrating RC-network; (b) at the output of a single-stage tuned amplifier? The excitation is white noise in either case. Are the realizations differentiable in the statistical sense?
3. What are the salient features of the autocorrelation function of the random signal at the output of a two-section RC-network excited by white noise?
4. When may we replace the real process existing at the output of a real circuit with white noise?
5. What is the noise bandwidth of a circuit?
6. List the physical factors that cause the output signal of a linear circuit to be normalized.
7. Describe the mechanism by which thermal noise arises in resistors. Define the frequency range within which thermal noise may be equated to white noise.
8. What determines the noise voltage between the terminals of a receiving antenna?
9. Describe the nature of shot noise in electron devices.
10. Write the equation defining the Poisson distribution and interpret the physical significance of the parameter ν.
11. What is the equivalent circuit used to represent the noise properties of an electron device displaying the shot effect?
12. How is the power spectrum of shot noise distributed in frequency?

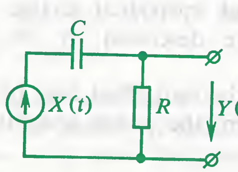
Problems

1. A stationary random process $X(t)$ whose power spectrum is

$$W_x(\omega)=\begin{cases}0, & \omega<-\omega_c\\ W_0, & -\omega_c<\omega<\omega_c\\ 0, & \omega>\omega_c\end{cases}$$

is applied to the input of an integrating RC-network. Find the variance of the output signal.

2. Work Problem 1 for the same signal applied to the input of the differentiating network shown in the accompanying diagram:



3. There is an ideal low-pass filter whose frequency response is

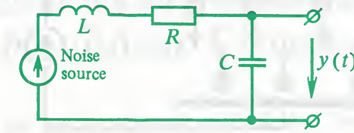
$$K(j\omega)=\begin{cases}0, & \omega<-\omega_c\\ K_0, & -\omega_c<\omega<\omega_c\\ 0, & \omega>\omega_c\end{cases}$$

where ω_c is its cut-off frequency. It is driven by a voltage which is a stationary random process whose autocorrelation function is

$$K_x(\tau)=\sigma_x^2\exp(-\alpha|\tau|)$$

Find the power spectrum, the autocorrelation function and the variance of the output voltage.

4. A series resonant circuit is driven by a white noise source whose power spectrum at all frequencies is W_0. Determine the power spectrum and the autocorrelation function of the output voltage $y(t)$.

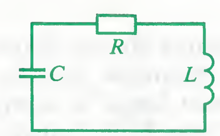


5. There is a single-stage tuned amplifier driven by a voltage source which generates Gaussian white noise with a power spectrum $F_0=10^{-15}\ \mathrm{V^2\ Hz^{-1}}$. The amplifier parameters are $K_{res}=100$, $Q_{eq}=110$, and $f_{res}=15$ MHz. Find the probability of the output noise voltage exceeding 2 mV.

6. Derive the equation defining the power spectrum of the output signal from a two-section RC-filter (see Example 10.3) driven by a noise voltage whose autocorrelation function is

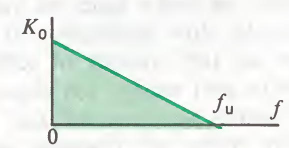
$$K_x(\tau)=\sigma_x^2\exp(-\alpha|\tau|)$$

7. The resonant circuit shown in the accompanying diagram



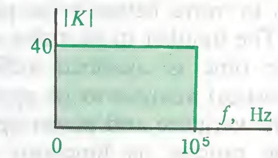
has the following parameters: $R=8\ \Omega$, $L=1.5\ \mu$H, and $C=120$ pF. It is placed inside an enclosure with a temperature $T=400$ K. Find the variance of the noise voltage across the inductor and the noise bandwidth of the circuit.

8. The amplitude response of a low-pass filter falls off linearly with rising frequency in the interval $(0, f_c)$: Find the noise bandwidth of the system.



9. Determine the noise bandwidth of the two-section RC-filter in Example 10.3.

10. A resistor $R=10$ kΩ held at a temperature $T=300$ K is connected to the input of an ideal low-pass filter having the following frequency response:



Determine the variance and the autocorrelation function of the filter's output signal.

11. A microwave antenna with a radiation resistance of $R_\Sigma=2.5\ \Omega$ picks up signals from a region in space at a temperature of $T=20$ K. The bandwidth of the system is 400 MHz. What is the rms noise voltage between the terminals of the antenna?

12. Determine the variance of indications by a sluggish instrument measuring the current of a thermionic diode. The direct component of current is 0.3 mA, and the characteristic integration time (time constant) of the instrument is 2 s.

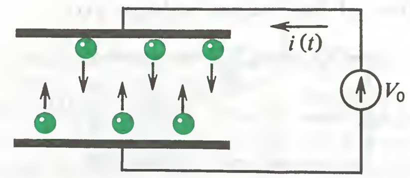
Advanced Problems

13. A random process $X(t)$ whose autocorrelation function is $K_x(t_1, t_2)$ is applied to the input of a linear stationary system for which the impulse response $h(t)$ is known. Using the Duhamel superposition integral, show by direct calculation that the autocorrelation function of the output signal is defined by

$$K_y(t_1, t_2)=\iint_{-\infty}^{\infty}K_x(\xi_1, \xi_2)\times h(t_1-\xi_1)h(t_2-\xi_2)\,d\xi_1 d\xi_2$$

that is, it is a two-dimensional convolution of $K_x(t_1, t_2)$ with $h(t)$.

14. Investigate the statistical characteristics of the random current in a system where electric-field forces cause a large number of minute conducting bodies (such as metal dust particles, small pieces of foil, and the like) to move between the plates of a capacitor. The number of particles moving at the same time is assumed sufficiently large for statistical analysis to be applicable. Determine the variance and power spectrum width of the process as functions of the system's physical parameters. Think up



a way of demonstrating the process. Propose its interpretation based on the shot noise model. Analyse the effect of the viscosity of the medium on the power spectrum width.

Chapter 11

Signal Transformations in Nonlinear Circuits

So far we have dealt with linear circuits. A salient feature of any linear circuit is that it obeys the principle of superposition. A simple and very important conclusion can be derived from that principle: On passing through a linear stationary system, a harmonic signal remains unchanged in waveform, but acquires a different amplitude and a different initial phase.

But it is precisely for this reason that a linear stationary system is not able to enrich the spectral composition of the waves applied to its input. This substantially limits the class of useful signal transformations that can be performed by constant-parameter linear circuits.

A wider choice in this respect is offered by nonlinear systems characterized by the fact that in them the input signal $v_{in}(t)$ and the response $v_{out}(t)$ are connected by a nonlinear functional relation of the form:

$$v_{out}(t) = f(v_{in},\ t) \tag{11.1}$$

In this chapter we will examine some general features inherent in nonlinear systems, techniques of mathematical analysis, and the most important of signal transformations that are performed by nonlinear circuits and devices.

11.1 Lag-Free (Zero-Memory) Nonlinear Transformations

In the general case, an analysis of a nonlinear circuit is a very complex task because in deriving a mathematical description for the internal state of the system we inevitably run into the need to solve nonlinear differential equations. It is a well-known fact that they cannot be solved by the techniques that yield a relatively easy solution for linear differential equations with constant coefficients. Yet, there are cases where an analysis of a nonlinear system can be carried to completion with relatively simple techniques. For this it will suffice to require that the nonlinear relation of the form in (11.1) *should not contain time explicitly*. Physically, this requirement implies that the nonlinear element involved is free from time lag (or has *zero memory*). In other words, its response instantaneously follows any changes in the excitation.

The condition for a transformation to be a lag-free (zero-memory) one

Strictly speaking, zero-memory nonlinear elements are nonexistent. However, this idealization is sufficiently accurate, if the

characteristic time of change in the input signal is substantially greater than the transient response time of the nonlinear element itself.

With regard to electronic circuits, nonlinear elements are most frequently semiconductor diodes and transistors. For their operation they depend on the diffusion of minority carriers in the regions of a semiconductor material directly adjacent to a *P-N* junction. Present-day semiconductor devices show a very high performance in terms of frequency properties. In them, the equilibrium (steady) state can be reached in a matter of 10^{-11} s. Therefore, the assumption that the internal processes in nonlinear circuit components are free from time lag is frequently well justified.

External characteristics of lag-free (zero-memory) nonlinear elements. The functional relation in (11.1) may be looked upon as the simplest mathematical model of a nonlinear element. Its distinction is that is does not involve any of the internal processes occurring in the element. It is customary to say that here we have to do with the external characteristic of a system.

To make matters more specific, the discussion that follows will be concerned with the external characteristics of nonlinear two-terminal networks (one-ports), assuming that the excitation is the voltage v across the one-port, and the response is the current i in that same one-port. The relation $i(v)$ is usually called the *current-voltage characteristic* of a nonlinear element. All the techniques involved and the results thus obtained may be extended to include four-terminal networks, or two-ports, such as transistors operating in the nonlinear region at large amplitudes of the input signal. Here, the output circuit is represented as a current source controlled by the output voltage; the relation between the instantaneous values of voltage and current is substantially nonlinear.

The current-voltage characteristic

Practical nonlinear elements have a variety of external characteristics. For example, there is a class of nonlinear elements having univalued current-voltage characteristics (Fig. 11.1*a*), and a class of nonlinear elements whose characteristics include multivalued regions (Fig. 11.1*b*).

The resistance of a nonlinear one-port. With regard to a nonlinear one-port the concept of resistance may be defined in more than one way. Let $i(v)$ be the current-voltage characteristic. On setting $v = V_0$, the current in the circuit will be $I_0 = i(V_0)$. The ratio

$$R_= = V_0/I_0 \tag{11.2}$$

is called the *d.c. resistance of the element*. In contrast to the usual resistance of a linear resistor, $R_=$ is not constant, but varies with the applied voltage.

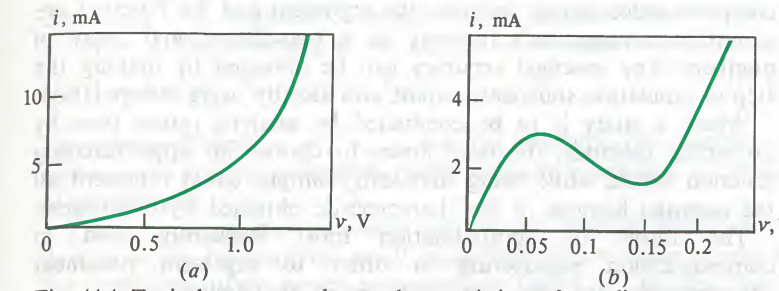
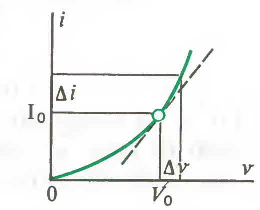


Fig. 11.1 Typical current-voltage characteristics of nonlinear two-port (four-terminal) networks: (*a*) single-valued characteristic of a crystal diode; (*b*) characteristic of a tunnel diode such that the same current can correspond to three different values of voltage

Frequently, a nonlinear element is driven by two voltage sources at the same time, V_0 and Δv, such that

$$|\Delta v|/|V_0| \ll 1$$



On expanding the current-voltage characteristic into a Taylor series about point V_0, we find the current to be

$$i \approx I_0 + i'(V_0)\Delta v$$

The ratio of a change in voltage to the resultant change in current at the selected operating point (V_0, I_0) is called the *dynamic* or *incremental resistance* of a nonlinear one-port

$$R_d = \Delta v/\Delta i = 1/i'(V_0) \tag{11.3}$$

Dynamic (incremental) resistance and transconductance

Sometimes it is more convenient to use the *dynamic transconductance*

$$g'_m = 1/R_d = di/dv\,|_{v=V_0} \tag{11.4}$$

defined by the slope of the current-voltage characteristic at the specified operating point.

It is to be stressed that the resort to the concept of dynamic resistance or dynamic transconductance is in effect the linearization of the actual current-voltage characteristic. It is valid solely for small increments in the signal about the operating point.

Description of the characteristic of nonlinear elements. As a rule, the current-voltage characteristics of nonlinear elements are obtained experimentally; they can be found theoretically far more seldom. For an analytic study of the processes occurring in circuits containing such elements, it is necessary above all to present the current-voltage characteristics in a mathematical form suitable for calculations.

The presentation of a characteristic in tabular form may be

a simple and very accurate method. It is especially convenient for computer-aided circuit analysis; the argument and the function are stored in a computer's memory as a two-dimensional array of numbers. Any specified accuracy can be obtained by making the step of tabulation sufficiently small, and also by using interpolation.

Where a study is to be conducted by analytic rather than by numerical methods, the need arises to choose an approximating function which, while being sufficiently simple, could represent all the essential features of the characteristic obtained by experiment.

The forms of approximation most frequently used in communication engineering in order to represent nonlinear characteristics are discussed in the pages that follow.

Piecewise-linear approximation. This technique is based on replacing a real characteristic curve by straight-line segments varying in slope. It is usually employed to analyse processes in nonlinear elements in the case of large-amplitude excitations. As an example, Fig. 11.2 shows the input characteristic of a real transistor approximated with two straight-line segments.

The approximation is determined by two parameters, namely the cut-off voltage V_c and the transconductance g_m which is the slope of the tangent to the curve at the specified points; its dimensions are units of conductance. Mathematically this is written as follows:

$$i(v) = \begin{cases} 0, & v < V_c \\ g_m(v - V_c), & v > V_c \end{cases} \qquad (11.5)$$

It is assumed that base current is substantially smaller than collector current

The cut-off voltage for the input characteristics of bipolar transistors is of the order of 0.2-0.8 V; the transconductance is usually around 10 mA/V. When instead of base current the dependent variable is collector current, with the base voltage being the independent variable, the value of transconductance must be multiplied by h_{21E}, the base current transfer ratio (or gain). Since $h_{21E} = 100\text{-}200$, the transconductance in this case is of the order of several amperes per volt.

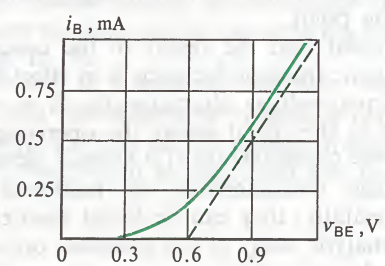


Fig. 11.2 Input characteristic of a transistor connected in a common-emitter circuit: base current is plotted as a function of base-emitter voltage

Power approximation. This form of approximation is based on expanding the nonlinear current-voltage characteristic $i(v)$ into a Taylor series converging in the vicinity of the operating point V_0:

$$i(v) = a_0 + a_1(v - V_0) + a_2(v - V_0)^2 + \ldots \qquad (11.6)$$

The general formula:

$$a_n = \frac{1}{n!} \left. \frac{d^n i}{dv^n} \right|_{v = V_0}$$

Here the coefficients $a_0, a_1, a_2, \ldots$ are real numbers. The number of terms in the expansion depends on the desired or specified accuracy of calculations.

Power approximation is widely used in the analysis of nonlinear devices driven by relatively small excitations. The manner in which the coefficients of the power expansion are found will be clear from the simple example that follows.

Example 11.1. *The experimental input characteristic $i_B = f(v_{BE})$ of a transistor is specified by the plot of Fig. 11.3. Find the coefficients a_0, a_1, and a_2 defining the approximation of the form*

$$i_B = a_0 + a_1(v_{BE} - V_0) + a_2(v_{BE} - V_0)^2$$

in the vicinity of the operating point $V_0 = 0.7$ V.

Let the nodes of the approximation be the points 0.5 V, 0.7 V, and 0.9 V. As is seen from the construction, in order to find the unknown coefficients, we should solve the following set of equations:

$$a_0 - 0.2a_1 + 0.04a_2 = 0.05$$

Solve Problem 1

$$a_0 = 0.15$$

$$a_0 + 0.2a_1 + 0.04a_2 = 0.5$$

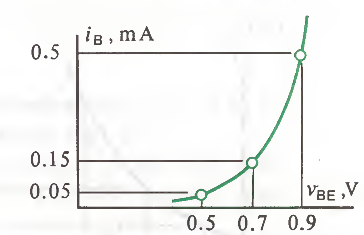


Fig. 11.3 Power approximation of the input characteristic of a transistor connected in a common-emitter circuit

Hence,

$$a_0 = 0.15 \text{ mA}, \quad a_1 = 1.125 \text{ mA/V}, \quad a_2 = 3.125 \text{ ma/V}^2$$

It is to be stressed that power approximation gives

predominantly a local description of a characteristic. It is not advisable to use this technique when the instantaneous value of the input signal markedly deviates from the operating point, as the accuracy of calculation would then be too low.

Exponential approximation. From the theory of *P-N* junctions, the following form is established for the current-voltage characteristic of a crystal diode for $v > 0$ near the origin

$$i(v) = i_0 [\exp(v/v_T) - 1] \tag{11.7}$$

Here i_0 is the leakage current across the junction, and v_T is the thermal voltage which is set at 25 mV for silicon devices as measured at a standard temperature of 300 K.

The exponential relation of the form (11.7) is frequently used in the study of nonlinear phenomena in electronic circuits containing semiconductor devices. The approximation is quite accurate for currents not exceeding several milliamperes. For heavier currents the exponential curve quietly changes into a straight line owing to the effect of the bulk resistance of the semiconductor material.

11.2 The Spectral Composition of the Current in a Zero-Memory Nonlinear Element Driven by a Harmonic Excitation

Consider what happens when a source of harmonic signal voltage

$$v_s(t) = V_m \cos \omega t$$

and a source of d.c. bias voltage V_0 drive a zero-memory (lag-free) nonlinear element. The waveform of the current in the circuit can be found by taking advantage of the simple graphic construction in Fig. 11.4.

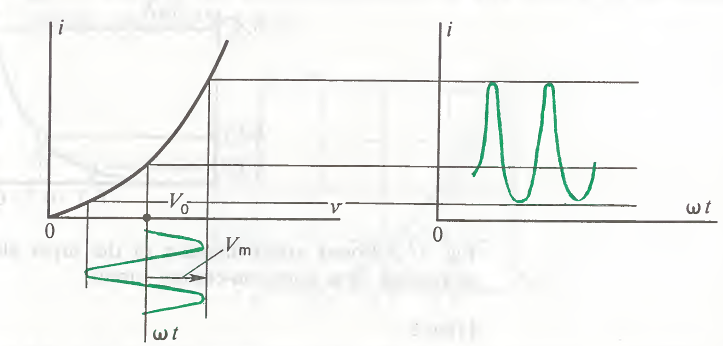


Fig. 11.4 Construction of a curve representing variations in the current in a lag-free nonlinear circuit

It is easy to see that the current and the voltage differ in waveform. The cause of distortion in the current waveform is very simple: *Equal changes in the voltage bring about unequal changes in the current because*

$$\Delta i = g_d(v) \Delta v$$

Also, the dynamic transconductance is likewise different within different portions of the curve.

The basic principle. If we approach the problem analytically, we will note that the function

$$i(t) = i(V_0 + V_m \cos \omega t)$$

defining the instantaneous values of current is a periodic one with period $T = 2\pi/\omega$. Therefore, it can always be expanded in a Fourier series

$$i(t) = I_0 + \sum_{n=1}^{\infty} I_n \cos(n\omega t + \varphi_n) \tag{11.8}$$

The spectral composition of current in a lag-free (zero-memory) one-port

Physically, this implies that the current in a lag-free nonlinear element is the sum of the d.c. component and, generally, an infinite set of harmonics at frequencies ω, 2ω, 3ω, ...

In engineering calculations, the most important task is to find the amplitudes of the spectral components of current (I_0, I_1, I_2, ...) as functions of the bias voltage and of the amplitude of the driving voltage V_m. Solution is carried out differently, according to the form of the approximating function used.

Piecewise-linear approximation. The waveform of current in a circuit containing a nonlinear element such that

$$i = \begin{cases} g_m(v - v_c), & v > V_c \\ 0, & v < V_c \end{cases} \tag{11.9}$$

where V_c is the cut-off (or cut-in) voltage, and driven by $v = V_0 + V_m \cos \omega t$, is seen from the construction in Fig. 11.5.

The current waveform is a succession of cosine pulses with cut-off. The spectral composition of such a periodic process was examined in detail in Chap. 2.

The cut-off angle* of current pulses can be found from the equation

$$V_0 + V_m \cos \vartheta = V_c$$

* The cut-off angle is half the angle of current flow or operating angle. – *Translators's note.*

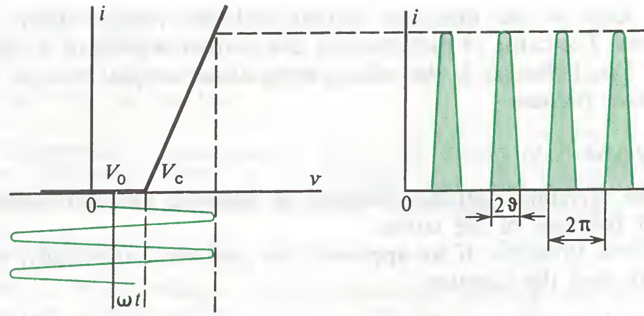


Fig. 11.5 The current in a circuit containing a piecewise-linear element

Hence,

$$\cos\vartheta = \frac{V_c - V_0}{V_m} \tag{11.10}$$

The direct component and the amplitudes of current harmonics are found by the equations

$$\begin{cases} I_0 = g_m V_m \gamma_0(\vartheta) \\ I_n = g_m V_m \gamma_n(\vartheta) \end{cases} \tag{11.11}$$

which include the corresponding Berg functions $\gamma_n(\vartheta)$.

Example 11.2. *A nonlinear element has a piecewise-linear current-voltage characteristics with the following parameters:* $V_c = 0.6$ V *and* $g_m = 25$ mA/V. *The voltage applied to the element is* $v = 0.2 + 0.8\cos\omega t$ (V). *Find the d.c. component* I_0 *and the fundamental* I_1 *of the current.*

Since

$$\cos\vartheta = \frac{0.6 - 0.2}{0.8} = 0.5$$

then

$$\vartheta = 60°$$

The values of the Berg functions are as follows:

$$\gamma_0 = \frac{1}{\pi}(\sin\vartheta - \vartheta\cos\vartheta) = 0.109$$

$$\gamma_1 = \frac{1}{\pi}(\vartheta - \sin\vartheta\cos\vartheta) = 0.196$$

Using Eqs. (11.11), we obtain:

$$I_0 = 25 \times 0.8 \times 0.109 = 2.18 \text{ mA}$$

$$I_1 = 25 \times 0.8 \times 0.196 = 3.92 \text{ mA}$$

Power approximation. Let

$$i = a_0 + a_1(v - V_0) + a_2(v - V_0)^2 + \dots \tag{11.12}$$

and the voltage applied to a nonlinear one-port be $v(t) = V_0 + V_m\cos\omega t$

Taking advantage of the well-known formulae

$$\cos^2 x = {}^1/_2(1 + \cos 2x)$$

$$\cos^3 x = {}^1/_4(3\cos x + \cos 3x)$$

$$\cos^4 x = {}^1/_8(3 + 4\cos 2x + \cos 4x)$$

$$\cos^5 x = {}^1/_{16}(10\cos x + 5\cos 3x + \cos 5x)$$

. . . .

we can re-write Eq. (11.12) as

$$\begin{aligned} i = {} & (a_0 + {}^1/_2 a_2 V_m^2 + {}^3/_8 a_4 V_m^4 + \dots) \\ & + (a_1 V_m + {}^3/_4 a_3 V_m^3 + {}^5/_8 a_5 V_m^5 + \dots)\cos\omega t \\ & + ({}^1/_2 a_2 V_m^2 + {}^1/_8 a_4 V_m^4 + \dots)\cos 2\omega t \\ & + ({}^1/_4 a_3 V_m^3 + {}^5/_{16} a_5 V_m^5 + \dots)\cos 3\omega t + \dots \end{aligned} \tag{11.13}$$

Hence, the following relations can be written for the direct component of current and the amplitudes of harmonics:

$$\begin{aligned} I_0 &= a_0 + {}^1/_2 a_2 V_m^2 + {}^3/_8 a_4 V_m^4 + \dots \\ I_1 &= a_1 V_m + {}^3/_4 a_3 V_m^3 + {}^5/_8 a_5 V_m^5 + \dots \\ I_2 &= {}^1/_2 a_2 V_m^2 + {}^1/_8 a_4 V_m^4 + \dots \\ I_3 &= {}^1/_4 a_3 V_m^3 + {}^5/_{16} a_5 V_m^5 + \dots \end{aligned} \tag{11.14}$$

It is to be noted that the d.c. component and the amplitudes of even harmonics are determined by the even-numbered coefficients of the Taylor series, and the odd harmonics depend solely on the odd-numbered coefficients.

Exponential approximation. If the current-voltage characteristic

of a nonlinear one-port can be approximated with an expression of the form

$$i(v) = i_0 [\exp(av) - 1]$$

the current spectrum can be found on the basis of the equation

$$\exp(x \cos \omega t) = I_0(x) + 2 \sum_{n=1}^{\infty} I_n(x) \cos n\omega t \tag{11.15}$$

where $I_n(x)$ is a modified Bessel function of index n.

If a nonlinear one-port with an exponential current-voltage characteristic is excited by the sum of the bias voltage and the harmonic signal voltage, that is

$$v = V_0 + V_m \cos \omega t$$

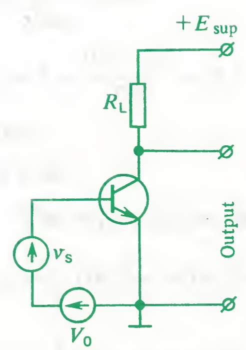
then

$$i(t) = i_0 [\exp(aV_0) I_0(aV_m) - 1]$$

$$+ 2i_0 \exp(aV_0) \sum_{n=1}^{\infty} I_n(aV_m) \cos n\omega t \tag{11.16}$$

Nonlinear distortion in a resistively loaded amplifier. The transformation that the spectrum of the input signal undergoes in nonlinear circuits is an extremely important matter. On the one hand, this spectrum transformation forms the basis for the operation of modulators, detectors and a number of other similar devices, which will be discussed below. On the other, the nonlinear behaviour gives rise to undesirable effects which must be properly evaluated and taken into consideration.

As an example of nonlinearity effects, we will examine a single-stage transistor amplifier loaded into a resistor, R_L. In contrast to a small-signal amplifier (see Chap. 8), it will be assumed that the amplitude of the input harmonic signal $V_{m,\text{in}}$ is sufficiently large to make it essential to consider the nonlinearity of the transfer characteristic $i_C = f(v_{BE})$ of the transistor connected in a common-emitter circuit. In the simplest case and with a properly positioned operating point, let this characteristic be specified by a second-order polynomial:



$$i_C = a_0 + a_1(v_{BE} - V_0) + a_2(v_{BE} - V_0)^2$$

If the amplifier is driven by the voltage

$$v_{BE} = V_0 + V_{m,\text{in}} \cos \omega t$$

the current in the collector lead will contain a steady (d.c.) component and also the fundamental and the second harmonic of

the signal, such that, on the basis of Eq. (11.14)

$$I_1 = a_1 V_{m,\text{in}}, \quad I_2 = {}^1/_2 a_2 V_{m,\text{in}}^2$$

The nonlinear distortion factor

These harmonic currents produce across the load resistor a voltage drop which is included in the output signal of the amplifier and distort it. Quantitatively, the degree of distortion in the output signal of an amplifier is stated in terms of the *nonlinear distortion factor* k_{nl} defined as the ratio of the rms value of all the higher harmonics (except the fundamental) to the amplitude of the useful signal:

$$k_{nl} = \frac{\sqrt{I_2^2 + I_3^2 + I_4^2 + \ldots}}{I_1} \tag{11.17}$$

In our case,

$$k_{nl} = I_2/I_1 = {}^1/_2 (a_2/a_1) V_{m,\text{in}} \tag{11.18}$$

Importantly, the nonlinear distortion factor increases with increasing amplitude of the signal.

11.3 Nonlinear Tuned Amplifiers and Frequency Multipliers

The key element of almost any radio transmitter is an r.f. tuned power amplifier. Its distinction is the fact that it operates at very large amplitudes of the input voltage. Therefore it is mandatory to take into account the nonlinearity of the current-voltage characteristic of the active element involved, which may be a transistor (transistors) or a tube (tubes).

Operating principle of a nonlinear tuned amplifier. Consider a single-stage transistor amplifier (Fig. 11.6*a*) loaded into a parallel resonant circuit (usually called the tank). The amplifier is driven by a voltage $v_{\text{in}}(t) = V_0 + V_{m,\text{in}} \cos \omega t$, and the resonant circuit is tuned to the signal frequency: $\omega_{\text{res}} = \omega$.

Suppose that the transistor characteristic $i_C = f(v_{BE})$ is approximated with straight-line segments and refer to Fig. 11.6*b*. The collector current waveform consists of a series of cosine pulses with cutoff (class C waveform). The pulses have a complex spectral composition, but the leading role in the operation of the amplifier is only played by the fundamental component whose frequency is the same as the resonant frequency of the tuned (or tank) circuit. The impedance of the tank circuit at harmonic frequencies 2ω, 3ω, ... is so low that the higher harmonics do not practically contribute to the output signal.

The fundamental of the collector current produces at the output a useful signal whose amplitude is

$$V_{m,\text{out}} = I_1 R_{\text{res}} = g_m R_{\text{res}} V_{m,\text{in}} \gamma_1(\vartheta) \tag{11.19}$$

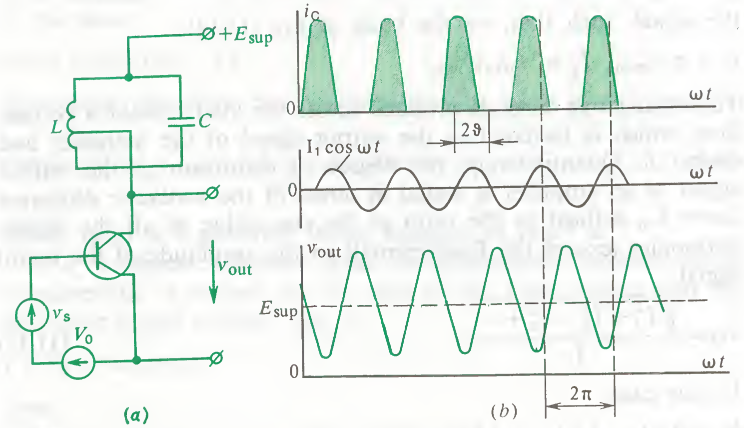


Fig. 11.6 Nonlinear tuned amplifier: (a) schematic diagram; (b) time waveforms of currents and voltages

The internal resistance of the source is absorbed in the resonant resistance

In a similar way, using Eq. (11.14) we may write an expression for the amplitude of the harmonic signal at the output of the tuned amplifier when power approximation is used for the transistor characteristic:

$$V_{m,out} = R_{res}(a_1 V_{m,in} + {}^3/_4\, a_3 V^3_{m,in} + {}^5/_8\, a_5 V^5_{m,in} + \ldots) \qquad (11.20)$$

The dynamic transfer characteristic of an amplifier. This refers to the relation $V_{m,out} = f(V_{m,in})$ stemming from Eq. (11.19) or (11.20). Naturally, it is required that the dynamic transfer characteristic be linear, especially in the case of amplifiers for AM signals. As follows from, say, Eq. (11.19), the dynamic transfer characteristic is in the general case nonlinear, because the cut-off angle ϑ and, as a consequence, the Berg function $\gamma_1(\vartheta)$ depend on the amplitude of the driving voltage $V_{m,in}$. The only exception is when the operating point is located at the origin. As can readily be seen, the cut-off angle is then $\vartheta = 90°$, irrespective of the value of $V_{m,in}$.

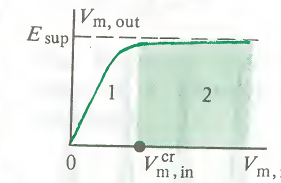
A further advantage of operation at a cut-off angle of 90° is that in the "no-signal" condition the d.c. component of collector current is zero. This factor has a wholesome effect on the efficiency of the amplifier.

An important parameter of the dynamic transfer characteristic is the width of its linear portion, as it determines the dynamic range of the signals being amplified. A natural factor limiting the rate of rise of the dynamic transfer characteristic is this: At some critical value of the input-signal amplitude, $V^{cr}_{m,in}$, the tuned-circuit voltage becomes comparable in magnitude with the supply voltage E_{sup}. Any further increase in the tuned-circuit voltage becomes impossible because at some instants of time the collector voltage of

▲ **Solve Problem 2**

the transistor crosses zero. In consequence, the normally turned-off collector junction is rendered conducting, and the tuned circuit of the amplifier is heavily shunted by the circuit containing the collector, base, signal source and power supply.

If $V_{m,in} > V^{cr}_{m,in}$, the amplifier is said to be *overdriven*. This condition cannot be used to amplify AM signals. However, by bringing down the supply voltage, a tuned amplifier can be brought to the overdriven condition, thereby turning it into an *amplitude limiter*, a device which suppresses the parasitic modulation of FM or PM signals.



(1) Underdriven condition; (2) overdriven condition

Power relations in a nonlinear tuned amplifier. Amplifiers of the type considered here handle a fairly large amount of power, so high efficiency is an important requirement for them. In order to determine the efficiency of an amplifier, we need to know the power drawn from the power supply

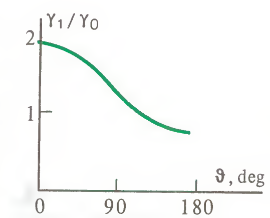
$$P_{sup} = I_0 E_{sup}$$

and the useful active power in the tuned circuit

$$P_{useful} = {}^1/_2\, I_1 V_{m,out}$$

In power amplifiers, it is sought to utilize the power supply to the utmost, even by coming very nearly to the overdriven operating condition, such that $V_{m,out} \approx E_{sup}$. Then

$$\text{efficiency} = P_{useful}/P_{sup} = {}^1/_2\gamma_1(\vartheta)/\gamma_0(\vartheta) \qquad (11.21)$$



If we analyse the ratio $\gamma_1(\vartheta)/\gamma_0(\vartheta)$, we will readily see that it is a maximum and equal to two for $\vartheta = 0°$. As ϑ is increased, the ratio falls off, being $\pi/2 = 1.571$ for 90°. Therefore, from the view-point of power utilization, it would seem advantageous to operate the amplifier at small values of ϑ, when the efficiency of the amplifier approaches unity. This would happen because the electron device would be turned off for the greater part of the cycle, and no power would be dissipated as heat at the collector (anode). This would, however, entail a drastic reduction in the value of γ_1, and in order to obtain the desired power output, we would have to substantially raise the amplitude of the input signal, which is not always possible. Therefore, bearing in mind the linearity of the transfer characteristic, the usual practice is to sacrifice some of the efficiency and to set the cut-off angle close to 90° (an operating angle of about 180°).

● **The choice of the cut-off angle (class of operation) for an amplifier**

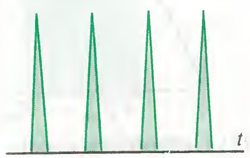
Resonance frequency multiplication. If we tune the tank circuit of a tuned amplifier operating at high amplitudes of the input signal to $n\omega$, which is the frequency of one of the signal's higher harmonics, the device can be used as a *frequency multiplier*.

A frequency multiplier comes in useful when one needs a highly stable source of harmonic waves which cannot, for one reason or

another, be generated directly, while a sufficiently stable low-frequency oscillator is available.

Often frequency multipliers are used in FM or PM transmitters in order to increase the frequency deviation of the signal. If at the input to a multiplier the frequency deviation is $\Delta\omega$, then at the multiplier's output it will obviously be $n\Delta\omega$, where n is the multiplier factor.

In principle, the design of a frequency multiplier does not differ from that of a nonlinear tuned amplifier. By analogy with Eq. (11.19), the amplitude of the output signal from a frequency multiplier in the case of piecewise-linear approximation is defined by



Output pulses for small values of the cut-off angle

$$V_{m,\text{out}} = g_m R_{\text{res}} V_{m,\text{in}} \gamma_n(\vartheta) \tag{11.22}$$

One of the difficulties arising in the synthesis of tuned frequency multipliers is that the function $\gamma_n(\vartheta)$ has a very low value at high values of the multiplier factor. Therefore, the cut-off angle must be chosen so as to maximize the corresponding Berg function. The smaller the ratio of the period to the duration of the collector current pulses*, the richer is their spectral composition. Hence, if the objective is to build a frequency multiplier with a high multiplier factor, the cut-off angle must be chosen very small. From analysis of the function $\gamma_n(\vartheta)$ it is seen that there exists an optimal cut-off angle, ϑ_{opt}, defined as

$$\vartheta_{\text{opt}} = 180°/n \tag{11.23}$$

This is the value of cut-off angle for a frequency multiplier operating at a fixed amplitude of the driving voltage $V_{m,\text{in}}$.

11.4 Lag-Free (Zero-Memory) Nonlinear Transformations of a Sum of Harmonic Signals

The property of a nonlinear network to enrich the spectrum of the output signal by adding to it the spectral components nonexisting in the input wave is most pronounced when the input signal is a sum of several harmonic waves differing in frequency. The appearance of a large number of new spectral components is the basis of the important nonlinear signal transformations that will be discussed in the next section.

The response of a nonlinear element having a power-approximation characteristic to a biharmonic excitation. We set out to elucidate the response of a nonlinear one-port whose current-voltage characteristic is described by, say, a second-order

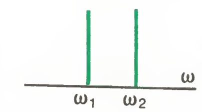
* This is the reciprocal of the *pulse duty factor*, τ/T.– *Translator's note.*

polynomial:

$$i(v) = a_0 + a_1(v - V_0) + a_2(v - V_0)^2 \tag{11.24}$$

In addition to the constant term V_0, the applied voltage contains two harmonic waves at different frequencies ω_1 and ω_2. Their amplitudes are V_{m1} and V_{m2} respectively:

$$v = V_0 + V_{m1}\cos\omega_1 t + V_{m2}\cos\omega_2 t \tag{11.25}$$



The frequency spectrum of a biharmonic signal

This is known as the *biharmonic excitation*. It is especially convenient in analysing spectrum transformations in nonlinear circuits.

On substituting the signal defined by (11.25) in Eq. (11.24), we obtain

$$\begin{aligned} i(t) = {} & a_0 + a_1 V_{m1}\cos\omega_1 t + a_1 V_{m2}\cos\omega_2 t \\ & + a_2 V_{m1}^2 \cos^2\omega_1 t + 2a_2 V_{m1} V_{m2}\cos\omega_1 t\cos\omega_2 t \\ & + a_2 V_{m2}^2\cos^2\omega_2 t \end{aligned}$$

After simple rearrangement we obtain the following spectral representation of the current in a nonlinear one-port:

$$\begin{aligned} i(t) = {} & \left[a_0 + \frac{a_2}{2}(V_{m1}^2 + V_{m2}^2)\right] + a_1 V_{m1}\cos\omega_1 t \\ & + a_1 V_{m2}\cos\omega_2 t + \frac{a_2 V_{m1}^2}{2}\cos 2\omega_1 t \\ & + \frac{a_2 V_{m2}^2}{2}\cos 2\omega_2 t + a_2 V_{m1} V_{m2}\cos(\omega_1 + \omega_2)t \\ & + a_2 V_{m1} V_{m2}\cos(\omega_1 - \omega_2)t \end{aligned} \tag{11.26}$$

As is seen, the current waveform contains the d.c. components, the fundamentals, and the second harmonics of the two input signals – they all were present prior to the transformation. In addition, it contains two new harmonic components at the sum and difference frequencies $\omega_1 + \omega_2$ and $\omega_1 - \omega_2$. Importantly, the amplitudes of these two waves, equal to $a_2 V_{m1} V_{m2}$, depend on the amplitudes of the two input signals to the same extent, and vanish when one of the two signals is not applied to the input. This is an indication that *the nonlinearity of the one-port is responsible for the interaction of the various harmonic components of the aggregate input signal.* A complete spectral diagram of the current in the one-port under consideration for the form of excitation chosen is shown in Fig. 11.7.

▲ Solve Problem 3

● Interaction of harmonics

The effect of the cubic term of the current-voltage characteristic.

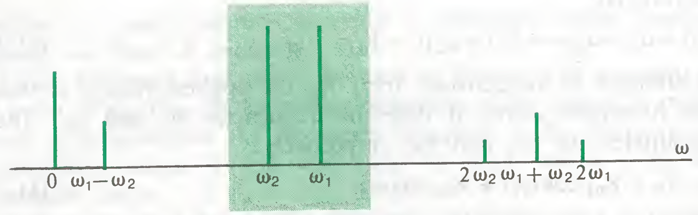


Fig. 11.7 Spectral diagram of the current in a nonlinear one-port whose current-voltage characteristic is described by a quadratic polynomial. (The input signal is a biharmonic wave whose spectrum is shown as a coloured block in the figure)

Let us make our problem somewhat more elaborate by assuming that the current-voltage characteristic $i(v)$ contains a cubic term responsible for an additional current defined by

$$i_3 = a_3(v - V_0)^3 \tag{11.27}$$

On inserting the signal defined by Eq. (11.25), we get

$$\begin{aligned} i_3(t) = a_3 [& (^3/_4 V_{m1}^3 + {}^3/_2 V_{m1} V_{m2}^2) \cos \omega_1 t \\ & + (^3/_4 V_{m2}^3 + {}^3/_2 V_{m1}^2 V_{m2}) \cos \omega_2 t \\ & + \frac{V_{m1}^3}{4} \cos 3\omega_1 t + \frac{V_{m2}^3}{4} \cos 3\omega_2 t \\ & + {}^3/_4 V_{m1}^2 V_{m2} \cos (2\omega_1 + \omega_2) t \\ & + {}^3/_4 V_{m1}^2 V_{m2} \cos (2\omega_1 - \omega_2) t \\ & + {}^3/_4 V_{m1} V_{m2}^2 \cos (2\omega_2 + \omega_1) t \\ & + {}^3/_4 V_{m1} V_{m2}^2 \cos (2\omega_2 - \omega_1) t] \end{aligned} \tag{11.28}$$

As is seen, the cubic term somewhat affects the amplitudes of the fundamental terms of current at frequencies ω_1 and ω_2. More importantly, however, there appear new spectral components at frequencies $3\omega_1$, $3\omega_2$, $2\omega_1+\omega_2$, $2\omega_1-\omega_2$, $2\omega_2+\omega_1$, and $2\omega_2-\omega_1$. Intermodulation frequencies. The example we have just analysed is a very special case because, firstly, the applied voltage has only two spectral components and, secondly, the current-voltage characteristic containing a quadratic or a cubic nonlinearity is very simple in form. The general statement of the problem is this: The current-voltage characteristic $i(v)$ is arbitrary; the input signal is represented by a trigonometric sum:

$$v(t) = V_0 + \sum_{k=1}^{\infty} V_{m,k} \cos \omega_k t \tag{11.29}$$

It is required to find the amplitudes and frequencies of all the spectral components present in the current.

This problem can be solved in a variety of ways, depending on the form of approximation used for the current-voltage characteristic [37]. The task of finding the amplitudes of the harmonic components usually involves cumbersome calculations, and we will not undertake to tackle it. In analysing the frequencies of the spectral components in the output signal, we can note, as has been confirmed by the previous example, that these so-called *intermodulation* (or *combination*) *frequencies* are specified by a general expression of the form

$$\omega = |n_1\omega_1 + n_2\omega_2 + \ldots + n_m\omega_m + \ldots| \tag{11.30}$$

where the n_i's are any whole numbers, positive and negative, including zero.

It is usual to group intermodulation frequencies by collecting the frequencies for which

$$|n_1| + |n_2| + \ldots + |n_m| + \ldots = N \tag{11.31}$$

Intermodulation frequencies

The number N is called the *order or degree of an intermodulation frequency.*

As the above example demonstrates, when a system is driven by a sum of two harmonic signals, the spectrum of the current flowing through the nonlinear element whose current-voltage characteristic contains terms of at most the third power includes intermodulation frequencies which can be grouped as follows:

N	Frequencies
1	ω_1, ω_2
2	$2\omega_1$, $2\omega_2$, $\omega_1+\omega_2$, $\omega_1-\omega_2$
3	$3\omega_1$, $3\omega_2$, $2\omega_1+\omega_2$, $2\omega_1-\omega_2$, $2\omega_2+\omega_1$, $2\omega_2-\omega_1$

It is important to note that the term of power N in the current-voltage characteristic is responsible for the appearance of intermodulation components whose highest order is equal to the power of that term. Also, if N is an even number, the associated intermodulation frequencies are likewise even: N, $N-2$, $N-4$, up to $N=0$ (the d.c. component). If, on the other hand, N is odd, the intermodulation frequencies are odd, too: N, $N-2$, $N-4$, up to $N=1$.

Work Problem 3

Example 11.3. *A nonlinear element has a cubic current-voltage characteristic*

$$i(v) = a_3(v - V_0)^3$$

The input voltage contains three harmonic waves:

$v = V_0 + V_{m1}\cos\omega_1 t + V_{m2}\cos\omega_2 t + V_{m3}\cos\omega_3 t$

Find the frequencies of all the intermodulation components of the current.

Since the power of the characteristic is three, there will be intermodulation frequencies with $N = 1$ and $N = 3$. The first-order intermodulation frequencies are ω_1, ω_2, and ω_3.

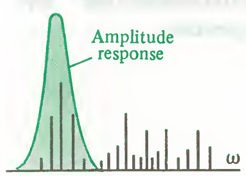
The third-order intermodulation frequencies are:

$3\omega_1,\ 3\omega_2,\ 3\omega_3,\quad |\pm\omega_1 \pm\omega_2 \pm\omega_3|$

$|\pm 2\omega_1 \pm \omega_2|,\quad |\pm 2\omega_1 \pm \omega_3|,\quad |\pm 2\omega_2 \pm \omega_1|$

$|\pm 2\omega_2 \pm \omega_3|,\quad |\pm 2\omega_3 \pm \omega_1|,\quad |\pm 2\omega_3 \pm \omega_2|$

Actually, only distinct frequencies need to be taken into consideration. The terms $2\omega_1 + \omega_2$ and $-2\omega_1 - \omega_2$ have the same frequency.



The filtering of intermodulation frequencies. The use of nonlinear devices for signal transformation is based on the following principle. When a lag-free (zero-memory) nonlinear element is driven by a sum of original waves, the output signal contains all likely combinations of intermodulation components. If now the output signal is passed through a linear frequency-selective network (a filter), a whole range of useful effects can be achieved because the output signal will contain spectral components previously nonexistent at the input.

One of the devices operating by this principle is the nonlinear tuned frequency multiplier that has already been examined. The spectrum transformation performed by a frequency multiplier is very simple because there is only one harmonic input signal.

11.5 Amplitude Modulation. Detection of AM Signals

The term "amplitude modulator" refers to a device which produces across its output terminals an AM signal of the form

$v_{AM}(t) = V_m(1 + M\cos\Omega t)\cos\omega_0 t$

when its input accepts the harmonic carrier wave

$v_{can}(t) = V_{m,car}\cos\omega_0 t$

and the baseband modulating signal

$v_{mod}(t) = V_{m,mod}\cos\Omega t$

Most often, amplitude modulators are built to utilize the transforma-

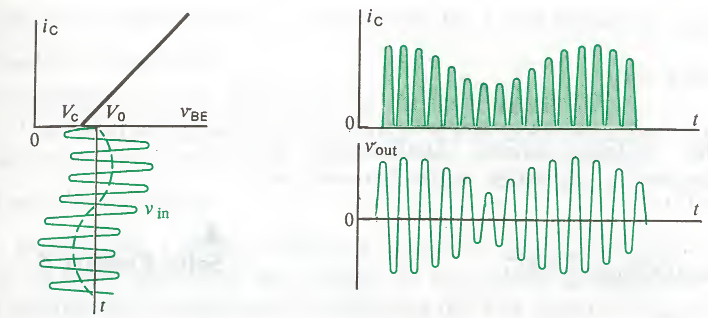
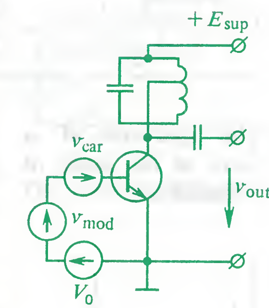


Fig. 11.8 Currents and voltages in an amplitude modulator

tion that the sum of the two applied signals undergoes in a lag-free (zero-memory) nonlinear element.



The principle of operation of an amplitude modulator. The simplest amplitude modulator is a single-stage nonlinear device loaded into a resonant (tuned) circuit. The input to the stage is a voltage of the form

$$v_{in}(t) = V_0 + V_{m,mod}\cos\Omega t + V_{m,car}\cos\omega_0 t \tag{11.32}$$

The resonant circuit in the collector lead is tuned to the carrier frequency. The operation of this amplitude modulator is explained for the current and voltage waveforms appearing in Fig. 11.8.

For definiteness, it is assumed that the current-voltage characteristic of the transistor is approximated with two straight-line segments. Since the Q-point moves in step with the baseband modulating wave, the cut-off angle of the carrier is continuously changing. Because of this the fundamental of the collector current pulse sequence is varying in time. The tuned circuit filters the collector current so that its output delivers an AM signal, that is, a wave whose amplitude varies in proportion to the modulating signal.

Example 11.4. *The current-voltage characteristic of the transistor used in a modulator circuit has a kink at $V_C = 0.6$ V. The amplitude of the carrier at the input is $V_{m,car} = 0.4$ V. The amplitude of the modulating signal is $V_{m,mod} = 0.1$ V. The quiescent bias voltage is $V_0 = 0.6$ V. Find the modulation factor M for the circuit in question.*

According to the statement of the problem, the Q-point moves between the limits $V_0 + V_{m,mod} = 0.7$ V and $V_0 - V_{m,mod} = 0.5$ V. Hence, the cut-off angle varies between

$$\vartheta_{max} = \arccos\frac{0.6 - 0.7}{0.4} = 1.823 \text{ rad}$$

and

$$\vartheta_{min} = \arccos\frac{0.6-0.5}{0.4} = 1.318 \text{ rad}$$

The amplitude of the collector-current fundamental is proportional to the Berg function $\gamma_1(\vartheta)$ which varies between the limits

Solve Problem 4

$$\gamma_1(\vartheta_{max}) = \frac{1}{\pi}(\vartheta_{max} - \sin\vartheta_{max}\cos\vartheta_{max}) = 0.657$$

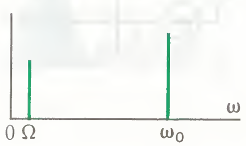
and

$$\gamma_1(\vartheta_{min}) = \frac{1}{\pi}(\vartheta_{min} - \sin\vartheta_{min}\cos\vartheta_{min}) = 0.342$$

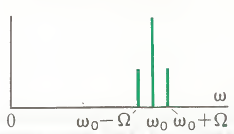
Hence, the modulation depth is

$$M = \frac{I_{1,max} - I_{1,min}}{I_{1,max} + I_{1,min}} = \frac{0.657-0.342}{0.657+0.342} = 0.315$$

The spectrum of a wave at the input of a modulator



and at its output



Analytical consideration. The production of an AM signal can be investigated analytically by using the above theory of intermodulation frequencies. Suppose that a nonlinear element having the simplest current-voltage characteristic of the form (11.24) is driven by the voltage

$$v(t) = V_0 + V_{m,car}\cos\omega_0 t + V_{m,mod}\cos\Omega t$$

such that $\omega_0 \gg \Omega$.

The current flowing through the one-port contains components whose frequencies are close to ω_0. These components make up the amplitude-modulated current

$$i_{AM}(t) = a_1 V_{m,car}\cos\omega_0 t + a_2 V_{m,car} V_{m,mod}\cos(\omega_0+\Omega)t + a_2 V_{m,car} V_{m,mod}\cos(\omega_0-\Omega)t \tag{11.33}$$

As will be recalled (see Chap. 4), the relative level of the side lobes in comparison with the carrier is $M/2$. From Eq. (11.33) it follows that in our case the modulation factor is

$$M = \frac{2a_2}{a_1}V_{m,mod} \tag{11.34}$$

Detection of AM signals. Detection (also called demodulation) refers to the process of separating the modulating signal from a modulated carrier. If the input signal to an ideal detector is an AM wave

$$v_{in}(t) = V_{m,in}(1 + M\cos\Omega t)\cos\omega_0 t$$

the wave appearing at its output will be a low-frequency signal

$$v_{out}(t) = V_{m,out}\cos\Omega t$$

proportional to the message signal being transmitted.

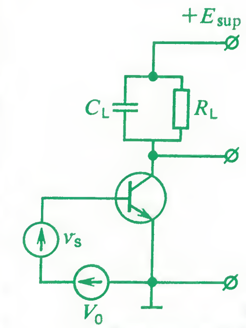
It is usual to define the performance of a detector in terms of the *detection* (or *demodulation*) *ratio*

The detection (demodulation) ratio

$$k_{det} = V_{m,out}/MV_{m,in} \tag{11.35}$$

Detection is a strictly nonlinear operation because the spectrum of the input wave does not contain the component at frequency Ω. Detection can be performed by applying the AM signal to a lag-free (zero-memory) nonlinear element and by filtering the resultant low-frequency components of the spectrum.

Consider the operation of what is called the *collector detector* which is a single-stage transistor circuit loaded into a parallel RC-network. For the load network to operate as a frequency filter suppressing all high-frequency spectral components, it is required that



The collector detector

$$1/\omega_0 C_L \ll R_L, \quad 1/\Omega C_L \gg R_L \tag{11.36}$$

This signifies that for the signal at the modulating frequency Ω the detector load is practically resistive and equal to R_L, whereas the magnitude of the load impedance and, in consequence the frequency response of the system at the carrier frequency ω_0 are negligibly small. Let

$$v_{in}(t) = V_0 + V_{m,in}(1 + M\cos\Omega t)\cos\omega_0 t$$

with $V_{m,in}$ being sufficiently large for piecewise-linear approximation to be applicable to the current-voltage characteristic of the nonlinear element. Also, to simplify matters, let $V_0 = V_c$ (which means that the system is biased at cut-off). Then the angle of cut-off will be $\vartheta = 90°$, irrespective of variations in the amplitude of the input signal in time. The operation of the collector detector is illustrated in the plots of Fig. 11.9.

Referring to Fig. 11.9, the collector current pulse sequence is amplitude-modulated; the zero current component is slowly varying (at frequency Ω), such that

$$I_{0,C} = g_m V_{m,in}(1 + M\cos\Omega t)\gamma_0(90°) = 0.318 g_m V_{m,in}(1 + M\cos\Omega t)$$

The output voltage of the detector is

$$v_{out}(t) = E_{sup} - I_{0,C}R_L = E_{sup} - 0.318 g_m R_L V_{m,in}(1 + M\cos\Omega t) \tag{11.37}$$

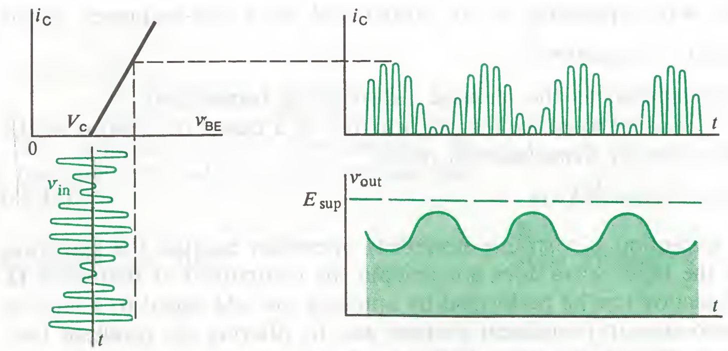


Fig. 11.9 Waveforms of currents and voltages in a collector detector

Hence, the detection ratio of the detector is

$$k_{det} = 0.318 g_m R_L \tag{11.38}$$

Linear detection

Importantly, the amplitudes of the input and output signals are in direct proportion. Quite aptly, this detector is said to effect *linear detection*. Its salient feature is freedom from distortion in the transmitted message.

Square-law detection. Let us consider separately a case of importance to applications, involving the detection of weak signals when the current-voltage characteristic of the nonlinear element has to be approximated with a power relation of the form

$$i_C(v) = a_0 + a_1(v_{in} - V_0) + a_2(v_{in} - V_0)^2 + \dots \tag{11.39}$$

We will limit ourselves to the terms written out and assume that the detector accepts an AM signal along with a d.c. bias voltage V_0:

$$v_{in}(t) = V_0 + V_{m,in}(1 + M\cos\Omega t)\cos\omega_0 t \tag{11.40}$$

On substituting (11.40) into (11.39), we find that among the various intermodulation waves present in the current, there is the following low-frequency component:

$$i_{lf}(t) = a_2 V^2_{m,in} M\cos\Omega t + \frac{a_2 V^2_{m,in}}{4} M^2\cos 2\Omega t \tag{11.41}$$

Owing to the filtering action of the load RC-network, the output signal will be determined by exactly this current:

$$v_{out}(t) = E_{sup} - a_2 R_L V^2_{m,in} M\cos\Omega t - \frac{a_2 R_L V^2_{m,in} M^2}{4}\cos 2\Omega t \tag{11.42}$$

Here the output voltage is proportional to the square of $V_{m,in}$, for

which reason this form of operation is called *square-law detection*. The presence of the term proportional to $\cos 2\Omega t$ in (11.42) is an indication that square-law detection is accompanied by distortion of the received message. On introducing the nonlinear distortion factor k_{nl}, equal to the ratio of the amplitudes of output waves at frequencies 2Ω and Ω, we find from Eq. (11.42) that $k_{nl} = M/4$. Hence it is clear that nonlinear distortion will be especially heavy at a high degree of modulation of the input signal. Therefore, in receivers it is desirable that the amplitude of the carrier in the AM signal applied to the detector's input be limited to a few volts. Then the detector will operate in the linear mode and there will be no nonlinear distortion in the received message.

Nonlinear distortion due to detection

Interaction of the valid signal and noise in a detector. Suppose that in addition to the valid AM signal the input of a square-law detector accepts an unmodulated noise wave whose frequency ω_n is close to the carrier frequency ω_0, such that

$$v_{in}(t) = V_0 + V_{m,in}(1 + M\cos\Omega t)\cos\omega_0 t + V_{m,n}\cos\omega_n t$$

The current due to the quadratic term in (11.39) is then equal to

$$i_2(t) = a_2(v_{in} - V_0)^2 = a_2 V^2_{m,in}(1 + M\cos\Omega t)^2\cos^2\omega_0 t$$
$$+ 2a_2 V_{m,in} V_{m,n}(1 + M\cos\Omega t)\cos\omega_0 t\cos\omega_n t + a_2 V^2_{m,n}\cos^2\omega_n t \tag{11.43}$$

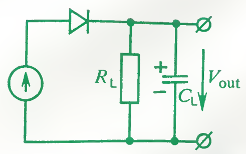
On the basis of Eq. (11.43), the additional low-frequency current component, which owes its origin to the noise wave, is equal to

$$\Delta i_{lf}(t) = {}^1/_2\, a_2 V_{m,in} V_{m,n} M[\cos(\omega_0 - \omega_n + \Omega)t$$
$$+ \cos(\omega_0 - \omega_n - \Omega)t] + a_2 V_{m,in} V_{m,n}\cos(\omega_0 - \omega_n)t$$

Thus, even with no modulation ($M = 0$), there is, at the detector's output, a low-frequency harmonic signal at frequency $\omega_0 - \omega_n$. When $M \neq 0$, the spectrum of the output signal becomes even more complex owing to the appearance of the low-frequency intermodulation waves at frequencies $\omega_0 - \omega_n + \Omega$ and $\omega_0 - \omega_n - \Omega$. These frequencies are close to those of the useful signal and cannot in principle be removed by filtering. If $V_{m,n} \gg V_{m,in}$, noise may substantially exceed the valid signal. In this sense, one speaks of the *suppression of a weak useful signal by a strong noise*. In order to minimize signal suppression, every effort should be made to reduce as much as possible the noise level during detection.

The diode AM detector*. This is a widely used type of detector,

* It is otherwise referred to as a *linear* or an *envelope detector*.— *Translator's note.*



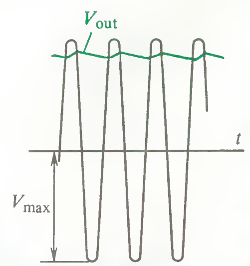
especially effective in large-signal operation. It consists of a series combination of a crystal diode and a parallel RC-network which acts as a frequency filter. The element values of the RC-network are chosen according to the condition defined in (11.36).

Assume that the diode has a piecewise-linear current-voltage characteristic and biased exactly at the cut-off voltage:

$$i(v) = \begin{cases} g_m v, & v > 0 \\ 0, & v < 0 \end{cases}$$

For the circuit to operate normally, it is essential that the load resistance R_L be substantially greater than the diode resistance in the forward direction, that is,

$$g_m R_L \gg 1$$



Let us apply to the detector input an unmodulated harmonic signal

$$v_{in}(t) = V_{m,in} \cos \omega_0 t$$

The capacitor charges via the forward resistance of the conducting diode much faster than it discharges via the high-value load resistor, and so the output signal is a sawtooth waveform with low teeth. The average level of the output voltage is close to the amplitude of the input signal. Thus, the diode remains nonconducting for a greater proportion of the cycle.

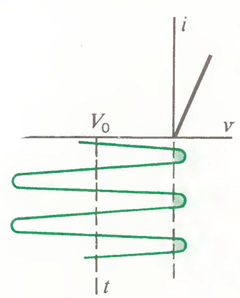
Let us neglect the minute variations in the output signal and take it that V_{out} is constant. We also note that V_{out} is applied to the diode in reverse polarity and acts as its bias voltage

$$V_0 = -V_{out}$$

The detection factor of the circuit

$$k_{det} = V_{out}/V_{m,in} = \cos \vartheta$$

may be close to unit, because the cut-off angle is sufficiently small.



The output voltage of a diode detector is close in amplitude to the input signal

The cut-off angle is found from the relation

$$-V_0 = I_0 R_L = g_m V_{m,in} \gamma_0(\vartheta) R_L$$

whence a transcendental equation follows

$$\cos \vartheta = \frac{g_m R_L}{\pi}(\sin \vartheta - \vartheta \cos \vartheta)$$

or

$$\tan \vartheta - \vartheta = \pi / g_m R_L \qquad (11.44)$$

For $g_m R_L \gg 1$, the cut-off angle is zero very nearly, so that the

following relation can be derived from Eq. (11.44) for the detection factor:

$$k_{det} = \cos(3\pi/g_m R_L)^{1/3} \qquad (11.45)$$

Work Problem 5

Example 11.5. *In a diode detector, $R_L = 18$ kΩ, and the diode transconductance is $g_m = 10$ mA/V. Find the detection factor.*

The product of the transconductance by the load resistance gives

$$g_m R_L = 180$$

Therefore, we may use Eq. (11.45), which yields

$$k_{det} = \cos(3.14 \times 3/180)^{1/3} = 0.93$$

When the input to a diode detector is an AM wave, then, subject to Eq. (11.36), the output voltage of the detector at any time is proportional to the instantaneous amplitude of the input signal (the modulated carrier).

11.6 Response of Lag-Free (Zero-Memory) Nonlinear Circuits to Stationary Random Signals

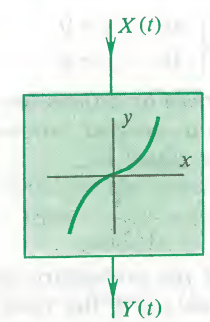
Suppose that the input to a lag-free (zero-memory) nonlinear system is a random signal $x(t)$ which is a realization of a stationary random process $X(t)$. The output signal $y(t)$ is connected to the input signal by a relation of the form

$$y(t) = f(x(t))$$

The ensemble of realizations $y(t)$ specifies a stationary random process $Y(t)$. We are to find the relation between the statistical characteristics of the processes $X(t)$ and $Y(t)$. This can be done, using any one of two approaches.

1. The nth-order multivariate probability density function of the input random process, $p(x_1, x_2, \ldots, x_n; t_1, \ldots, t_n)$, is known and we are to find a similar function, $p(y_1, y_2, \ldots, y_n; t_1, \ldots, t_n)$, for the output process.

2. Acting within the framework of correlation theory, we set out to find the mean m_y and the autocorrelation function $K_y(\tau)$ of the output random process. In addition to the autocorrelation function, it may be of interest to find the power spectrum $W_y(\omega)$ of the output signal.



The probability density of the output signal after nonlinear transformation. The first objective is achieved by using the techniques set forth in Chap. 6 for the probability densities of systems of random variables subjected to functional transformations. If x_1, x_2,

..., x_n are the random variables observed at the input at times t_1, t_2, ..., t_n, respectively, then, since the transformation is lag-free, we will have at the output and at the same instants of time the following variables

$$y_1 = f(x_1),\ y_2 = f(x_2),\ \ldots,\ y_n = f(x_n) \tag{11.46}$$

By using the inverse function $x = \varphi(y)$, we have

$$x_1 = \varphi(y_1),\ x_2 = \varphi(y_2),\ \ldots,\ x_n = \varphi(y_n) \tag{11.47}$$

The multivariate probability density for the output process is

$$p_{\text{out}}(y_1, \ldots, y_n) = p_{\text{in}}(\varphi(y_1), \ldots, \varphi(y_n))\,|D| \tag{11.48}$$

in which the Jacobian D has a very simple form

In the case of multivalued functions, a sum should be taken of the contributions from all branches

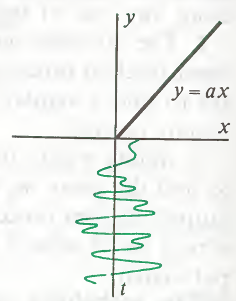
$$D = \begin{vmatrix} \dfrac{d\varphi(y_1)}{dy_1} & 0 & \cdots & 0 \\ 0 & \dfrac{d\varphi(y_2)}{dy_2} & & \vdots \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \dfrac{d\varphi(y_n)}{dy_n} \end{vmatrix}$$

$$= \frac{d\varphi}{dy}\bigg|_{y=y_1} \frac{d\varphi}{dy}\bigg|_{y=y_2} \cdots \frac{d\varphi}{dy}\bigg|_{y=y_n} \tag{11.49}$$

Equation (11.48) yields the solution in the most general form. It is to be noted that the structure of the Jacobian determinant (11.49) stems from the assumption that the transformation involved is a lag-free (zero-memory) one – the instantaneous value of the output signal solely depends on the input signal at the same instant of time.

Example 11.6. *A lag-free (zero-memory) nonlinear element having the piecewise-linear characteristic*

$$y = \begin{cases} ax, & x > 0 \\ 0, & x < 0 \end{cases} \tag{11.50}$$



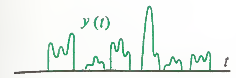
is driven by a Gaussian random process $X(t)$ of mean value zero and with a specified variance σ_x^2. The probability density function of the input signal is

$$p_{\text{in}}(x) = \frac{1}{\sqrt{2\pi}\sigma_x} \exp(-x^2/2\sigma_x^2)$$

Find the probability density function of the output signal.

For $x > 0$, the inverse function has the form

$x = y/a$

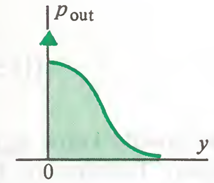


and so

$dx/dy = 1/a$

Therefore,

$$p_{\text{out}}(y) = \frac{1}{\sqrt{2\pi}\sigma_x a} \exp(-y^2/2\sigma_x^2 a^2)$$



for $y > 0$. For any negative value of x there is only one value $y = 0$. To assure the normalization of the probability density function for the output signal, we should allow a δ-singularity in the output probability density $p_{\text{out}}(y)$ for $y = 0$ with a coefficient equal to 1/2:

$$p_{\text{out}}(y) = \begin{cases} \dfrac{1}{2}\delta(y) + \dfrac{1}{\sqrt{2\pi}\sigma_x a} \exp(-y^2/2\sigma_x^2 a^2) & \text{for } y \geqslant 0 \\ 0 & \text{for } y < 0 \end{cases} \tag{11.51}$$

It is fundamentally important that, on applying a Gaussian signal to the input of a nonlinear system, we observe a non-Gaussian random process at its output.

The mean value of the signal at the output of a nonlinear system. The simplest statistical characteristic of a stationary random process is its mean obtained by averaging its realizations over the ensemble or, if the process is ergodic, over one sufficiently long realization. In order to be able to find the mean of a signal subjected to a nonlinear, lag-free (zero-memory) transformation, we should have at our disposal the univariate probability density function $p_{\text{out}}(y)$. Then

$$\bar{y} = m_y = \int_{-\infty}^{\infty} y p_{\text{out}}(y)\,dy \tag{11.52}$$

Thus, the problem reduces to finding a quadrature. With equal success, we could find the mean of a transformed signal by averaging the function $f(x)$ with the aid of the univariate probability density function of the input signal:

$$\bar{y} = \int_{-\infty}^{\infty} f(x)\,p_{\text{in}}(x)\,dx \tag{11.53}$$

Example 11.7. *Find the mean of the output signal for the system of Example 11.6.*

By Eq. (11.53),

$$\bar{y} = \frac{1}{\sqrt{2\pi}} \int_0^{\infty} (ax/\sigma_x) \exp(-x^2/2\sigma_x^2)\,dx = a\sigma_x/\sqrt{2\pi} = 0.399 a\sigma_x \tag{11.54}$$

The result implies that the variance of a stationary Gaussian process can be measured with the aid of a nonlinear converter having a characteristic of the form defined in (11.50), connected in cascade with a sluggish linear network which performs averaging over time.

The autocorrelation function of the output signal. By the general rule, the autocorrelation function of the signal $y(t)$ at the output of a lag-free nonlinear converter is

$$K_y(\tau) = \overline{y(t)\,y(t+\tau)} - \bar{y}^2$$

$$= \iint\limits_{-\infty}^{\infty} f(x) f(x_\tau)\, p_{\text{in}}(x, x_\tau;\ \tau)\, \mathrm{d}x\,\mathrm{d}x_\tau - \bar{y}^2 \tag{11.55}$$

So that we could make use of Eq. (11.55), we should know $p_{\text{in}}(x, x_\tau;\ \tau)$ which is the bivariate probability density function of the input signal observed at two sections separated by a time interval τ.

Calculations by Eq. (11.55) may prove fairly cumbersome. The final result in a more or less observable form can only be obtained for a normal input process when

$$p_{\text{in}}(x, x_\tau;\ \tau) = \frac{1}{2\pi\sigma_x^2(1-R_x^2)^{1/2}} \exp\left[-\frac{x^2 + x_\tau^2 - 2R_x x x_\tau}{2\sigma_x^2(1-R_x^2)}\right] \tag{11.56}$$

where $R_x(\tau)$ is the correlation coefficient for the input signal.

Example 11.8. *Find the autocorrelation function of the output signal under the conditions given in Example 11.6.*

The main difficulty is to find the mean product

$$\overline{yy_\tau} = \frac{a^2}{\pi} \int\limits_0^\infty \int\limits_0^\infty \frac{x x_\tau}{2\sigma_x^2(1-R_x^2)^{1/2}} \exp\left[-\frac{x^2 + x_\tau^2 - 2R_x x x_\tau}{2\sigma_x^2(1-R_x^2)}\right] \mathrm{d}x\,\mathrm{d}x_\tau$$

By a change of variables,

$$\xi = \frac{x}{\sigma_x\sqrt{2(1-R_x^2)}}, \quad \xi_\tau = \frac{x_\tau}{\sigma_x\sqrt{2(1-R_x^2)}}$$

we may write the mean product as

$$\overline{yy_\tau} = \frac{2}{\pi} a^2\sigma_x^2 (1-R_x^2)^{3/2} I$$

where

$$I = \int\limits_0^\infty \int\limits_0^\infty \xi\xi_\tau \exp(-\xi^2 - \xi_\tau^2 + 2R_x\xi\xi_\tau)\, \mathrm{d}\xi\,\mathrm{d}\xi_\tau$$

The simplest way to evaluate the last integral is by changing to polar coordinates:

$$\xi = \rho\cos\varphi, \quad \xi_\tau = \rho\sin\varphi$$

Omitting simple, but cumbersome calculations, the final result is

$$I = \frac{1}{4(1-R_x^2)^{3/2}}\left[\sqrt{1-R_x^2} + R_x \arccos(-R_x)\right]$$

Hence, using Eq. (11.54) we find the autocorrelation function of the output signal:

$$K_y(\tau) = \frac{a^2\sigma_x^2}{2\pi}\left[\sqrt{1-R_x^2} + R_x \arccos(-R_x) - 1\right] \tag{11.57}$$

Since for $\tau = 0$, the correlation coefficient is $R_x(0) = 1$, the variance of the output signal is

$$\sigma_y^2 = K_y(0) = \frac{a^2\sigma_x^2}{2\pi}(\pi - 1) = 0.3408 a^2\sigma_x^2 \tag{11.58}$$

Therefore, the correlation coefficient of the random process at the output of a piecewise-linear nonlinear converter is given by

Work Problems 6 and 7

$$R_y(\tau) = \frac{1}{\pi - 1}\left[\sqrt{1-R_x^2} + R_x \arccos(-R_x) - 1\right] \tag{11.59}$$

Nonlinear transformations of narrowband random processes. Suppose that the input to a lag-free (zero-memory) nonlinear converter is a Gaussian narrowband random process. Its realizations are quasiharmonic random waves with central frequency ω_0. The autocorrelation function of the input signal is

$$K_x(\tau) = \sigma_x^2 R_x(\tau) = \sigma_x^2 \rho(\tau) \cos\omega_0\tau \tag{11.60}$$

Let us find the autocorrelation function of the output signal with special reference to the piecewise-linear element examined in the previous examples. The direct substitution for R_x from (11.60) into (11.59) yields the desired result, but this technique is not so obvious. It is advantageous to rearrange Eq. (11.59) by expanding its right-hand side into an infinite series in powers of $R_x(\tau)$. To this end, we will take advantage of the fact that

$$\arccos(-R_x) = \pi/2 + R_x + R_x^3/6 + \ldots$$

and

$$\sqrt{1-R_x^2} = 1 - R_x^2/2 - R_x^4/8 - \ldots$$

Therefore,

$$R_y(\tau) = \frac{1}{\pi - 1}(\pi R_x/2 + R_x^2/2 + R_x^4/24 + \ldots) \tag{11.61}$$

On the basis of (11.60) we find that

$$R_y(\tau) = \frac{1}{\pi - 1}\left[\frac{\pi}{2}\rho(\tau)\cos\omega_0\tau + 1/2\rho^2(\tau)\right.$$
$$\left.\times \cos^2\omega_0\tau + 1/24\rho^4(\tau)\cos^4\omega_0\tau + \ldots\right] \quad (11.62)$$

The effect of the higher terms in the expansion is negligible

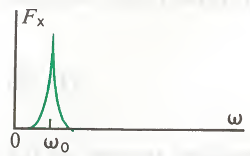
Since

$$\cos^2\omega_0\tau = 1/2 + 1/2\cos 2\omega_0\tau$$

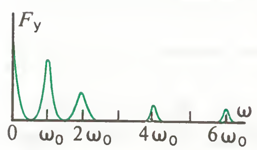
and

$$\cos^4\omega_0\tau = 3/8 + 1/2\cos 2\omega_0\tau + 1/8\cos 4\omega_0\tau$$

it follows from (11.62) that



$$R_y(\tau) \approx 0.117\rho^2(\tau) + 0.007\rho^4(\tau) + 0.733\rho(\tau)\cos\omega_0\tau$$
$$+ [0.117\rho^2(\tau) + 0.01\rho^4(\tau)]\cos 2\omega_0\tau$$
$$+ 0.002\rho^4(\tau)\cos 4\omega_0\tau \quad (11.63)$$



Hence, by taking advantage of the W-K theorem, we are able to identify the form of the power spectrum $F_y(\omega)$ of the output random process. As it turns out, the power spectrum of the wave at the output of the nonlinear converter in question breaks up into an infinite sum of components each of which represents an individual narrowband random process. The "masses" (peaks) of their power spectra are observed at frequencies ω_0, $2\omega_0$, $4\omega_0$, ... Also, the power spectrum of the output signal includes a low-frequency component in the vicinity of the zero signal, which may be looked upon as the outcome of the amplitude detection of the input signal.

It is interesting to note that in this special case the spectrum of the output signal does not contain components at frequencies $3\omega_0$, $5\omega_0$, ... Undoubtedly, should the nonlinear element have any other form of characteristic, we might expect the appearance of all harmonics of the central frequency of the input random wave. As a rule, the intensity of the high-frequency spectral components rapidly falls off with increasing harmonic number.

Summary

- The real current-voltage characteristics of lag-free (zero-memory) nonlinear one-ports can be approximated with a variety of simple functions. Most frequently, use is made of piecewise-linear, power and exponential approximations.
- The current flowing in a lag-free (zero-memory) one-port driven by a harmonic excitation contains in the general case a d.c. component and an infinite number of harmonics, that is, waves at frequencies which are multiples of the input-signal frequency.
- The output voltage of a tuned amplifier operating in the nonlinear mode is a sinewave owing to the frequency-selective behaviour of the tuned circuit, despite the nonharmonic nature of the current flowing through the tuned circuit.

- When the input signal has a large amplitude, the tuned amplifier is operating in the overdriven mode.
- The response of a nonlinear element to a sum of harmonic signals differing in frequency contains waves at intermodulation frequencies.
- The filtering of appropriate intermodulation waves makes it possible to effect both amplitude modulation and amplitude demodulation (detection).
- Under a lag-free (zero-memory) nonlinear transformation, a Gaussian random process gives rise to a non-Gaussian random process.
- In order to find the autocorrelation function of a random process at the output of a nonlinear converter, one has to have the bivariate probability density function of the input signal.

Review Questions

1. Define the condition under which the nonlinear transformation of a signal may be taken as being lag-free.
2. What current-voltage characteristic must a nonlinear element have for its output current to be free from odd harmonics?
3. Define the considerations governing the choice of the cut-off angle for a large-signal tuned amplifier.
4. State the physical principle underlying the operation of a nonlinear frequency multiplier. Why is it that a high multiplication factor is difficult to achieve?
5. Define the considerations governing the choice of the load parameters for an AM detector.
6. What is the difference between linear and square-law detection?
7. What determines the choice of the cut-off angle for a diode detector?
8. Describe the calculation of a univariate and a multivariate probability density function for a random process after it has been subjected to a lag-free (zero-memory) nonlinear transformation.
9. What is the salient feature of the nonlinear transformation of narrowband random processes?

Problems

1. A nonlinear one-port has a current-voltage characteristic (mA) of the form

$$i(v) = 10v^3$$

What is the analytic expression for this characteristic in the vicinity of the operating point $V_0 = 2$ V?

2. A harmonic tuned amplifier is set up as shown in Fig. 11.6. The characteristic of the transistor (mA) is approximated by two straight-line segments:

$$i_C(v_{BE}) = \begin{cases} 50(v_{BE} - 0.2) & \text{for } v_{BE} > 0.2\text{ V} \\ 0 & \text{for } v_{BE} < 0.2\text{ V} \end{cases}$$

The tuned-circuit impedance at resonance is $R_{res} = 0.8$ kΩ. The supply voltage is $E_{sup} = 9$ V. The operating point is positioned at the kink of the characteristic. Find the

amplitude of the input signal that will overdrive the amplifier.

3. A nonlinear lag-free (zero-memory) element has a current-voltage characteristic of the form $i(v) = a_0 + a_1 v + a_4 v^4$.

The voltage applied to the element is

$v(t) = V_{m1} \cos \omega_1 t + V_{m2} \cos \omega_2 t$

Find the amplitudes and frequencies of all the intermodulation components of current.

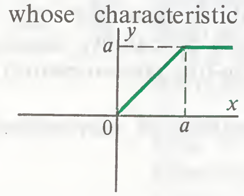
4. There is a nonlinear resistor whose characteristic (mA) has the form

$i(v) = 25v + 4v^2$

The voltage applied to the resistor is given (in V) by $v = 5 + 2\cos \Omega t + 1.5 \cos \omega_0 t$. Find the amplitude of the carrier and the depth of modulation.

5. There is a diode detector built around a crystal diode whose transconductance is $g_m = 10$ mA/V. The load resistor is $R_L = 20$ kΩ. The applied AM signal (in V) is given by $v(t) = 5(1 + 0.6 \cos \Omega t) \cos \omega_0 t$. Find the amplitude of the low-frequency (Ω) signal developed across the detector load.

6. There is a nonlinear device which is a limiter whose characteristic $y = f(x)$ is

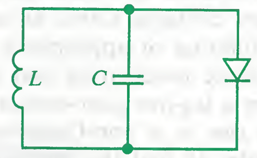


The input of the limiter accepts a Gaussian stationary random process $X(t)$ of mean value zero and of variance σ_x^2. Plot for the output realization. Find the mean and the variance of the output signal.

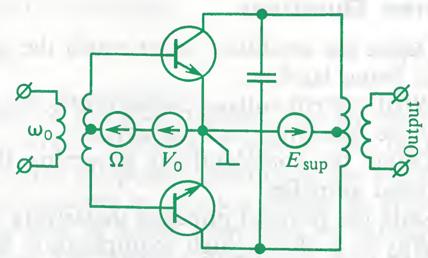
7. A lag-free (zero-memory) nonlinear element has a characteristic of the form $y = a|x|$ and is driven by a Gaussian stationary noise of mean value zero and with an autocorrelation function $K_x(\tau) = \sigma_x^2 R_x(\tau)$. Find the expression for the autocorrelation function of the output signal.

Advanced Problems

8. Analyse the free response of a nonlinear system which is a parallel LC tuned circuit shunted by a crystal diode.



9. Analyse the operation of the balanced modulator shown in the accompanying circuit diagram:



The tuned circuit in the collector lead is tuned to ω_0. Show that, if the circuit is completely symmetrical, the output signal will not contain a component at the carrier frequency.

10. Show that the diode detector set up as shown in the accompanying schematic

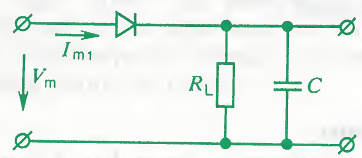


diagram has an input impedance of $R_{in} = R_L/2$. The input impedance is defined as the ratio of the amplitude V_m of the input voltage to the amplitude of the fundamental component I_{m1} of the input current.

11. Propose a circuit capable of demodulating angle-modulated signals. The circuit should include a linear frequency-selective network and an amplitude detector.

Chapter 12

Signal Transformations in Linear Parametric Circuits

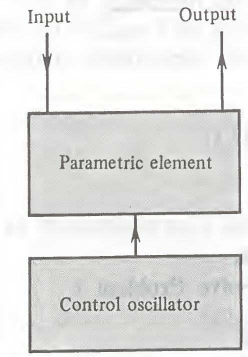
Linear systems described by time-dependent nonstationary system operators $T(t)$ possess a number of properties of interest and value for communication engineering applications. Here, the input signal is transformed as defined by

$$v_{out}(t) = T(t)\, v_{in}(t) \tag{12.1}$$

such that, owing to the linearity of the system,

$$T(t)(\alpha_1 v_{in1} + \alpha_2 v_{in2}) = \alpha_1 T(t)\, v_{in1} + \alpha_2 T(t)\, v_{in2} \tag{12.2}$$

for any constants α_1 and α_2.



Circuits described by Eq. (12.1) are called *parametric*. The name arises from the fact that such circuits always contain elements whose parameters are time-dependent. The parametric devices most commonly used in communication applications are resistors $R(t)$, capacitors $C(t)$, and inductors $L(t)$.

A salient feature of a linear parametric system is the inclusion of an auxiliary oscillator which controls the parameters of the circuit elements.

The prominence given to parametric circuits in telecommunications is due to the ability of such systems to transform the spectra of input signals, and also to the possibility of building low-noise parametric amplifiers.

12.1 Response of Resistive Parametric Circuits

The resistive parametric network

A parametric circuit is called *resistive (zero-memory)*, if its system operator is a number $k(t)$ which depends on time and serves as a coefficient of proportionality between the input signal $v_{in}(t)$ and the output signal $v_{out}(t)$:

$$v_{out}(t) = k(t)\, v_{in}(t) \tag{12.3}$$

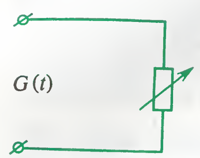
The simplest system of this type is a parametric resistor of resistance $R(t)$. The relation connecting the instantaneous values of voltage and current in this one-port is

$$v(t) = R(t)\, i(t) \tag{12.4}$$

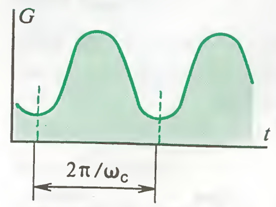
Alternatively, a resistive parametric element may be described in terms of the time-variant conductance

$G(t) = 1/R(t)$

The spectrum of the current in a resistive parametric one-port. As a rule, the control signal applied to the parametric element from an external auxiliary source is periodic in time. Therefore, in studying signal transformations in resistive parametric elements, we will concentrate on the controlled resistor whose conductance $G(t)$ is representable by a Fourier series:



$$G(t)=G_0/2+\sum_{k=1}^{\infty} G_k\cos(k\omega_c t-\Psi_k) \tag{12.5}$$



where ω_c is the frequency of the control oscillator, and G_0, G_1, G_2, ... and Ψ_1, Ψ_2, ... are constant numbers determining the amplitudes and phases of the corresponding harmonics in the conductance spectrum.

Let the resistive parametric one-port be driven by a harmonic signal voltage

$$v(t)=V_m\cos(\omega_s t+\varphi_s)$$

Then the current in the one-port will be

$$i(t)=G(t)v(t)=\frac{V_mG_0}{2}\cos(\omega_s t+\varphi_s)$$

$$+\sum_{k=1}^{\infty}V_mG_k\cos(k\omega_c t-\Psi_k)\cos(\omega_s t+\varphi_s)$$

$$=\frac{V_mG_0}{2}\cos(\omega_s t+\varphi_s)+\frac{V_m}{2}\sum_{k=1}^{\infty}G_k\cos[(\omega_s+k\omega_c)t$$

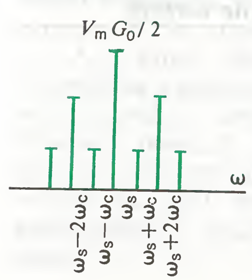
$$+\varphi_s-\Psi_k]+\frac{V_m}{2}\sum_{k=1}^{\infty}G_k\cos[(\omega_s-k\omega_c)t+\varphi_s+\Psi_k] \tag{12.6}$$

▲ **Solve Problem 1**

■ **The spectrum of current in a resistive parametric one-port**

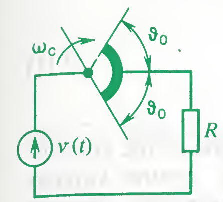
Equation (12.6) can be simplified by taking sums over the positive and negative indices k, noting that $G_k=G_{-k}$ *and* $\Psi_k=-\Psi_{-k}$, so that

$$i(t)={}^1/_2\,V_m\sum_{k=-\infty}^{+\infty}G_k\cos[(\omega_s+k\omega_c)t+\varphi_s-\Psi_k] \tag{12.7}$$



Equation (12.7) defines the spectral composition of the current in a resistive parametric one-port: This current contains a component at the signal frequency with an amplitude $V_mG_0/2$, and also a number of intermodulation waves at frequencies $\omega_s+k\omega_c$ ($k=\pm1,\pm2,\pm3,\ldots$) and with amplitudes $V_mG_k/2$. The spectral diagram of the current is symmetrical about ω_s.

It may be noted that in a sense the spectral composition of the current in a parametric resistor is simpler than that in a lag-free (zero-memory) nonlinear element. For one thing, it does not contain any harmonics of the signal frequency.

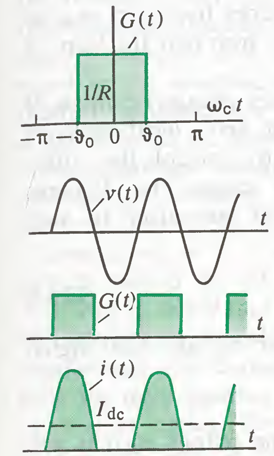


The parametric network with a resistor and a switch

Example 12.1. *Consider a parametric network formed by a fixed resistor R and a switch which connects the resistor periodically to a voltage source $v(t)=V_m\cos\omega_s t$. The period of the control signal is T_c. The switch operates so that during the time interval $(-T_c/2, T_c/2)$ the circuit is completed for all t's satisfying the condition $-\vartheta_0<\omega_c t<\vartheta_0$, and is open for the remaining instants of time. Find the amplitudes of all spectral components of the current.*

As follows from the statement of the problem, the conductance $G(t)$ of the network is an even function that can be expanded into a Fourier series of the form (12.5) only in terms of cosines with zero initial phases Ψ_k. The corresponding Fourier coefficients are

$$G_k=(2/\pi R)\int_{-\vartheta_0}^{\vartheta_0}\cos k\xi\,d\xi=(2/k\pi R)\sin k\vartheta_0 \tag{12.8}$$



The intermodulation current components at frequencies $\omega_s+k\omega_c$ will have the amplitudes

$$I_k=(V_m/k\pi R)\sin k\vartheta_0 \tag{12.9}$$

Interestingly, if the input signal and the control wave are synchronous ($\omega_s=\omega_c$), a steady (d.c.) component will be present

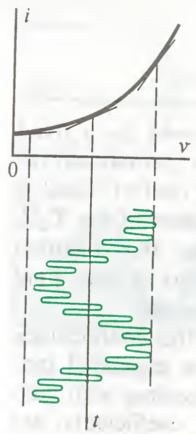
$$I_{dc}=(V_m/\pi R)\sin\vartheta_0,\quad k=-1$$

The effect of current rectification at $\theta=90°$

In the circumstances, the parametric network operates as a rectifier: the current is produced only by the positive half-cycles of the cosine wave.

Implementation of resistive parametric elements. In practice parametrically controlled resistors are implemented as follows. A lag-free (zero-memory) nonlinear one-port whose current-voltage characteristic is $i=f(v)$ is driven by a sum of two waves: the control voltage $v_c(t)$ and the signal voltage $v_s(t)$. The control voltage is chosen to be substantially greater in magnitude than the signal. The current thus produced in the nonlinear one-port may be written by expanding the current-voltage characteristic in a Taylor series in terms of the instantaneous value of control voltage:

$$i=f(v_c+v_s)=f(v_c)+f'(v_c)v_s+\frac{1}{2}f''(v_c)v_s^2+\ldots \tag{12.10}$$



The dynamic transconductance is determined by the “large” control voltage

The signal amplitude is chosen so small that we may neglect the second and higher powers of $v_s(t)$ in Eq. (12.10). By using $i_s(t)$ to designate the increment in the current through the one-port due to the presence of the signal, we get

$$i_s(t) \approx f'(v_c(t))\,v_s = g_d(v_c(t))\,v_s \tag{12.11}$$

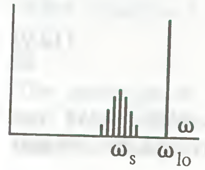
As is seen, the dynamic transconductance (the slope of the current-voltage characteristic) is determined by the “large” control voltage. The most important applications of resistive parametric elements will be discussed in the pages that follow.

The intermediate frequency

Frequency conversion. This refers to the transformation of the spectrum of a modulated signal, consisting in that the modulated signal is translated from the vicinity of the carrier frequency ω_{car} to the vicinity of some *intermediate frequency* ω_i, such that the form of modulation remains unaffected.

A *frequency converter* (also called a *frequency changer*) consists of a *mixer* which is a lag-free parametric element, and a *local oscillator* operating at frequency ω_{lo} which serves to control the mixer parametrically. The local-oscillator voltage causes the dynamic slope of the current-voltage characteristic of the mixer to vary periodically in a manner defined by

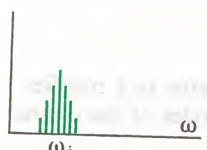
$$g_d(t) = g_0 + g_1 \cos \omega_{lo} t + g_2 \cos 2\omega_{lo} t + \ldots \tag{12.12}$$



If the input to the frequency converter is an AM signal

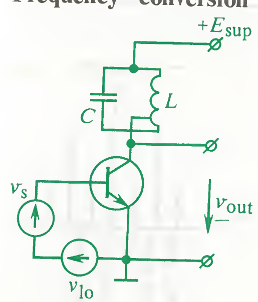
$$v_s(t) = V_m(1 + M \cos \Omega t) \cos \omega_s t$$

then, by virtue of Eqs. (12.11) and (12.12), the output current will contain a component defined by



Frequency conversion

$$i_s(t) = V_m(1 + M \cos \Omega t)[g_0 \cos \omega_s t + {}^1/_2\, g_1 \cos(\omega_{lo} - \omega_s)t$$
$$+ {}^1/_2\, g_1 \cos(\omega_{lo} + \omega_s)t + {}^1/_2\, g_2 \cos(2\omega_{lo} - \omega_s)t$$
$$+ {}^1/_2\, g_2 \cos(2\omega_{lo} + \omega_s)t \ldots]$$



Frequency converter

Customarily, the intermediate frequency (i.f.) is chosen such that $\omega_i = |\omega_{lo} - \omega_s|$. The current at the intermediate frequency

$$i_{if}(t) = \frac{g_1 V_m}{2}(1 + M \cos \Omega t) \cos \omega_i t \tag{12.13}$$

is an AM wave modulated in the same manner as the input signal.

In order to extract the spectral components at frequencies close to the intermediate frequency, the output circuit of the frequency converter includes a tuned circuit tuned to ω_i.

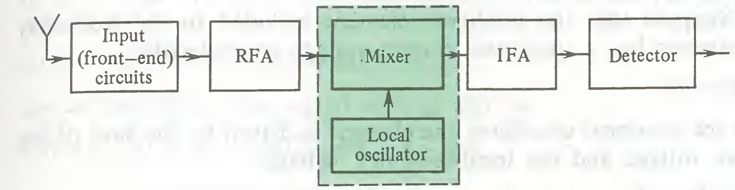


Fig. 12.1 Block diagram of a superheterodyne receiver

The operating principle of the superhet receiver

Frequency conversion is widely used in *superheterodyne receivers.* A block diagram of a superheterodyne (or superhet) receiver is given in Fig. 12.1. It operates as follows. The signal picked up by the antenna is conveyed via the input circuits and an r.f. amplifier (RFA) to the frequency converter. The output signal of the frequency converter is a modulated wave whose carrier frequency is equal to the intermediate frequency of the receiver. The bulk of receiver amplification and receiver selectivity is assured by a narrowband i.f. amplifier (IFA).

An important distinction of a superhet receiver is that its intermediate frequency remains unchanged. In tuning a superhet receiver, one only has to vary the tuning of the local oscillator and, in some cases, of the tuned circuits that may be present in the input circuits and in the RFA.

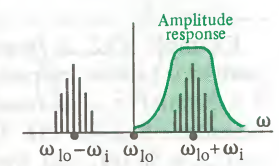


Image rejection

It is to be noted that the frequency converter responds equally to two frequencies, namely $\omega_{s1} = \omega_{lo} + \omega_i$ and $\omega_{s2} = \omega_{lo} - \omega_i$. In radio engineering, it is customary to say that reception is possible over both the main channel and its image channel. To avoid ambiguity in receiver tuning, the selectivity of the resonant circuits placed between the antenna and the frequency changer is adjusted in such a way that practically all of the image-frequency signal is rejected.

Conversion transconductance

Conversion transconductance. The effectiveness of a frequency converter is customarily stated in terms of a special parameter called the *conversion transconductance,* g_c. It is defined as the ratio of the amplitude of current at the intermediate frequency to the amplitude of the unmodulated signal voltage:

$$g_c = I_{m,if}/V_{m,s}$$

As follows from Eq. (12.13),

$$g_c = g_1/2 \tag{12.14}$$

Or, in words, the conversion transconductance is equal to half the fundamental of the dynamic transconductance of the parametric element.

Suppose that the nonlinear element included in the frequency converter has a quadratic current-voltage characteristic:

$i(v) = bv^2$

In the no-signal condition, the element is driven by the sum of the bias voltage and the local-oscillator voltage:

$v_c = V_0 + V_{m,lo} \cos \omega_{lo} t$

The dynamic transconductance of the frequency converter varies in time as given by

▲ **Work Problem 3**

$$g_d = 2bv_c = 2bV_0 + 2bV_{m,lo} \cos \omega_{lo} t \tag{12.15}$$

Referring to (12.14), we can see that in this case

$$g_c = bV_{m,lo} \tag{12.16}$$

Thus, with the input signal held constant, the amplitude of the output signal of the frequency converter is directly proportional to the amplitude of the local-oscillator voltage.

Let us illustrate the calculation of a frequency converter with a simple example.

Example 12.2. *A frequency converter uses a nonlinear element (a transistor) whose current-voltage characteristic is $i_C = 20v_{BE}^2$, so that $b = 20$ mA/V^2. The tuned circuit in the collector lead has a resonant resistance $R_{res} = 3$ kΩ. The amplitude of the unmodulated input signal is $V_{m,s} = 50$ μV, and the amplitude of the local-oscillator voltage is $V_{m,lo} = 0.5$ V. Find the amplitude of the i.f. voltage at the converter output, $V_{m,if}$.*

From Eq. (12.16), the conversion transconductance is

$g_c = 20 \times 0.5 = 10$ mA/V

The amplitude of the output current at the intermediate frequency is

$I_{m,if} = g_c V_{m,s} = 0.5$ μA

Assuming that the output impedance of the transistor is sufficiently high for its shunting effect on the tuned circuit to be negligible, the amplitude of the i.f. voltage is

$V_{m,if} = 1.5$ mV

Synchronous detection. Suppose that the local oscillator of a frequency converter is tuned exactly to the signal frequency, and, in consequence, the dynamic transconductance varies in time as given by

$g_d(t) = g_0 + g_1 \cos \omega_s t + g_2 \cos 2\omega_s t + ...$

If the voltage applied to the frequency converter is an AM signal defined by

$v_s(t) = V_{m,s}(1 + M \cos \Omega t) \cos (\omega_s t + \varphi_s)$

the current due to the signal voltage will be

$$i_s(t) = V_{m,s}(1 + M \cos \Omega t)[g_0 \cos (\omega_s t + \varphi_s) + {}^1/_2 g_1 \cos (2\omega_s t + \varphi_s) + {}^1/_2 g_1 \cos \varphi_s + ...] \tag{12.17}$$

Here the expression in the square brackets contains a steady (d.c.) component, $(g_1/2) \cos \varphi_s$, which depends on the phase shift between the local-oscillator voltage and the input-signal carrier. Therefore, the spectrum of the output current acquires a low-frequency component

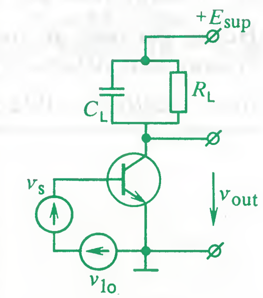
$$i_{lf}(t) = (V_{m,s} g_1/2)(1 + M \cos \Omega t) \cos \varphi_s \tag{12.18}$$

This current faithfully reproduces the message being transmitted.

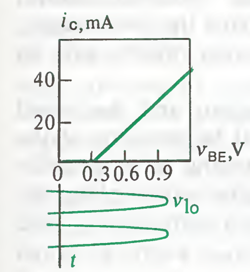
A *synchronous detector* (or *synchronous demodulator*) is a frequency converter operating subject to the condition $\omega_{lo} = \omega_s$. In order to extract the useful signal, the output circuit includes a low-pass filter, such as a parallel *RC*-network.

In practice, the use of a synchronous detector is handicapped by the fact that an exact phase relation must be maintained between the input-signal carrier and the local-oscillator voltage. The best choice is $\varphi_s = 0°$. If, on the other hand, $\varphi_s = 90°$, the useful output signal vanishes. However, this sensitivity of a synchronous detector to the phase shift can be utilized in measuring the phase relations between two coherent waves.

The example that follows illustrates the calculation of a specific type of synchronous detector.



The synchronous detector (demodulator)

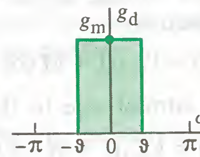


Example 12.3. *A synchronous detector is built around a transistor whose characteristic $i_C = f(v_{BE})$ is approximated with two straight-line segments. The approximation parameters are: $g_m = 50$ mA/V and $V_c = 0.3$ V. The amplitude of the local-oscillator voltage is $V_{m,lo} = 1$ V; the direct bias voltage is zero ($V_0 = 0$). The unmodulated signal voltage with an amplitude $V_{m,s} = 25$ μV is shifted in phase relative to the local oscillator through an angle $\varphi_s = 45°$. Find variations in the d.c. voltage at the output of the synchronous detector caused by the useful signal, if $R_L = 1.2$ kΩ.*

In the case of a piecewise-linear nonlinear element the dynamic transconductance (the slope of the current-voltage characteristic) may take on only two values:

$$g_d = \begin{cases} 0, & v_c < V_{cutoff} \\ g_m, & v_c > V_{cutoff} \end{cases}$$

Therefore variations in the dynamic transconductance with time can be represented by a periodic series of rectangular video pulses. The cut-off angle ϑ which defines the duration of the pulses can be found from (see Chap. 2)



$$\vartheta = \arc\cos\left(\frac{V_{cutoff} - V_0}{V_{m,lo}}\right) = 72.5°$$

In order to determine the amplitude of the fundamental term of the transconductance, we may use Eq. (12.8) in which the conductance $1/R$ is replaced with g_m, the transconductance of the transistor within the active region of the characteristic. Then

$$g_1 = (2g_m/\pi)\sin\vartheta = 0.607 g_m = 30.35 \text{ mA/V}$$

In accord with Eq. (12.18), the useful signal brings about the following increment in the current flowing through the transistor:

$$\Delta i = (V_{m,s} g_1/2)\cos\varphi_s = 0.268\ \mu\text{A}$$

Hence, variation in the direct voltage level at the output of the synchronous detector is

$$\Delta v = -\Delta i R_L = -0.32 \text{ mV}$$

12.2 Energy and Power Relations in Reactive Parametric Elements

Reactive parametric elements, that is, those in which either the capacitance, $C(t)$, or the inductance, $L(t)$, is varying with time, display a whole range of spectral properties. Taking a parametrically controlled capacitor as an example, we will demonstrate that, given certain conditions, such elements can act as "intermediaries" which transfer power from external control sources, called *pump oscillators*, to circuits carrying the useful signal. This principle underlies *parametric amplification* which will be studied in the next section.

The pump oscillator

Relation between the capacitance of a capacitor and the stored energy. To get proper insight into the physical foundations of the processes taking place in reactive parametric circuits, let us consider the following idealized system. Let a flat capacitor whose plates are spaced a distance x_0 apart, be charged to a voltage V_0. The capacitor carries a separated charge $Q = CV_0$, where C is its capacitance.

An increase in the spacing between the plates of a capacitor leads to a decrease in its capacitance

Suppose that by some mechanical means the capacitor plates are moved apart and the spacing between them is increased to $x_0 + dx$. The displacement is directed against the electric-field force which tends to move the capacitor plates closer together. In overcoming

the field force, external forces have to do some positive work on the system, with the result that the field energy stored in the capacitor is increased.

In order to obtain quantitative results, we note that the original energy stored in the capacitor is

$$E = Q^2/2C \tag{12.19}$$

When the capacitance is incremented by dC, the energy increment is

$$dE = -(Q^2/2C^2)\,dC = -E\,dC/C \tag{12.20}$$

because there is no conduction current flowing, and the charge Q remains unchanged. On finding the capacitance C from the equation for a flat capacitor (known from physics),

The electric constant

$$\varepsilon_0 = 10^{-9}/36\pi = 8.842 \times 10^{-12} \text{ F m}^{-1}$$

$$C = \varepsilon_0 A/x_0$$

(where A is the surface area of a capacitor plate), we obtain the following expression for the relative increment in capacitance:

$$dC/C = -dx/x_0$$

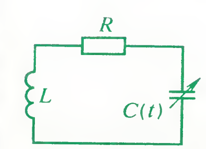
On inserting the above result in (12.20), we get

$$dE = E\,dx/x_0 \tag{12.21}$$

As should be expected, Eqs. (12.20) and (12.21) show that in order to increase the electric-field energy stored in a system, the capacitance of the charged capacitor must be decreased by some external factors.

Parametric excitation of a resonant circuit. So long as the capacitor charge Q is held constant, no continuous inflow of energy into the system can be brought about by varying the capacitor capacitance about some mean value. The point is that on doing a positive work over the time interval while the capacitance is decreased, the external source will receive from the capacitor the same amount of energy during the next time interval when the capacitance is increased. Naturally, the energy averaged over a period (or cycle) will be zero.

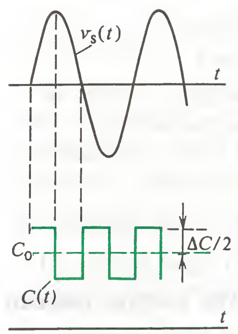
The picture is entirely different in a resonant circuit when the voltage across the capacitor due to the free process crosses zero and changes sign. Consider a high-Q resonant circuit formed by a fixed inductance L, a parametric capacitance $C(t)$, and a loss resistance R. Suppose that in some unspecified manner natural (free) oscillations have been produced in the tuned circuit. Neglecting the insignificant decrease in the amplitude of the oscillations due to losses, let us take it that $V_{m,C}$, the amplitude of the voltage across the capacitor, varies in time at the frequency of natural (free)



oscillations

$$\omega_{nat} = 1/\sqrt{LC_0}$$

where C_0 is the mean capacitance of the parametric capacitor.



Assume that the capacitance of the capacitor varies periodically in the following manner: Twice over a period of natural oscillations the capacitance undergoes a negative-going jump by ΔC at the instants when the voltage across the capacitor is at an extremum (positive or negative), and is restored to its original value or, which is the same, undergoes a positive-going jump by the same amount at the instants when the voltage across the capacitor crosses zero.

This pumping action will result in a unidirectional flow of energy into the resonant circuit. Indeed, the work done by external forces at the instants when a negative change of capacitance takes place will always be positive, irrespective of the sign of the voltage across the capacitor. On the other hand, no energy will be expended at the instants when the voltage across the capacitor is zero and there is no electric field to give rise to forces of repulsion.

If

$$E_{C,max} = \frac{1}{2} V_{m,C}^2 (C_0 + \Delta C/2) \approx \frac{1}{2} V_{m,C}^2 C_0$$

is the maximum energy stored in the capacitor, then, in accord with Eq. (12.20), the energy pumped into the tuned circuit over a period of natural oscillations will be

Note that energy is supplied to the tuned circuit twice every period of natural oscillations

$$E_{pump} = 2E_{C,max} \Delta C/C_0 = V_{m,C}^2 \Delta C \tag{12.22}$$

At the same time, the mean power lost in the tuned circuit will be

$$P_{loss} = \frac{1}{2} I_m^2 R = \frac{1}{2} \frac{V_{m,C}^2}{\rho^2} R = \frac{1}{2} \frac{V_{m,C}^2}{\rho Q} \tag{12.23}$$

The total power dissipated in the resistor over a period of oscillations, T, will be

$$E_{loss} = P_{loss} T = V_{m,C}^2 C_0 \pi/Q \tag{12.24}$$

It is taken that $\rho\omega_{nat} = 1/C_0$

If the equality

$$E_{pump} = E_{loss} \tag{12.25}$$

is satisfied, the pump source will make up for the losses in the tuned circuit. If, on the other hand, $E_{pump} > E_{loss}$, the system will turn unstable, and the amplitude of the oscillations in the circuit will build up exponentially – the oscillatory circuit is said to be *excited parametrically*. From Eqs. (12.22) and (12.24), we can derive the following relation defining the critical relative variation in capacitance:

$$\Delta C_{cr}/C_0 = \pi/Q \tag{12.26}$$

As a rule, the values of ΔC_{cr} are small. As an example, for parametric excitation of a tuned circuit with $C_0 = 20$ pF and $Q = 100$, it will suffice for ΔC to be 0.63 pF.

In the above analysis, we proceeded from the assumption that the pump signal changes the capacitance of the capacitor twice over a period of the natural oscillations. It is easy to see, however, that the parametric excitation of the oscillatory system will take place also when the fundamental frequency of the pump voltage is $\omega_{pump} = \omega_{nat}/n$ ($n = 1, 2, \ldots$). What is important is that the spectrum of the pump signal should contain a component at frequency $2\omega_{nat}$.

The requirements that the pump should meet in order to assure the self-excitation of the tuned circuit

It is also required that a precise phase relation be maintained between the natural oscillations of the tuned circuit and those of the pump oscillator. A shift of the pump signal by a half-cycle in phase so that the positive change of capacitance occurs at the instants when the voltage across the capacitor passes through an extremum will cause the parametric capacitor to act as an additional resistive load instead of as an energy source.

Relation between the voltage across and the current in a parametric capacitor. Consider a circuit formed by a source of signal voltage

$$v(t) = V_m \cos(\omega_s t + \varphi_s)$$

and a controlled capacitor whose capacitance varies in time harmonically at the pump frequency:

$$C(t) = C_0 [1 + \beta \cos(\omega_{pump} t + \varphi_{pump})] \tag{12.27}$$

Here β is a coefficient defining the degree of modulation of the capacitance. Since the charge on the capacitor is

$$q = C(t)\, v$$

the current in the circuit is

$$\begin{aligned} i(t) &= \frac{dC}{dt} v + C \frac{dv}{dt} \\ &= -\beta\omega_{pump} C_0 V_m \sin(\omega_{pump} t + \varphi_{pump}) \cos(\omega_s t + \varphi_s) \\ &\quad - \omega_s C_0 V_m \sin(\omega_s t + \varphi_s) \\ &\quad - \beta\omega_s C_0 V_m \cos(\omega_{pump} t + \varphi_{pump}) \sin(\omega_s t + \varphi_s) \end{aligned} \tag{12.28}$$

Taking advantage of a known trigonometric formula

$$\cos x \sin y = {}^1/_2 [\sin(x + y) - \sin(x - y)]$$

we may write the products in the first and third terms on the right-

hand side of Eq. (12.28) as follows:

$$\sin(\omega_{pump}t + \varphi_{pump})\cos(\omega_s t + \varphi_s)$$

$$= {}^1/_2 \{\sin[(\omega_s + \omega_{pump})t + \varphi_s + \varphi_{pump}]$$

$$- \sin[(\omega_s - \omega_{pump})t + \varphi_s - \varphi_{pump}]\} \quad (12.29)$$

▲ Work Problem 4

$$\cos(\omega_{pump}t + \varphi_{pump})\sin(\omega_s t + \varphi_s)$$

$$= {}^1/_2 \{\sin[(\omega_s + \omega_{pump})t + \varphi_s + \varphi_{pump}]$$

$$- \sin[(\omega_{pump} - \omega_s)t + \varphi_{pump} - \varphi_s]\} \quad (12.30)$$

Thus,

$$i(t) = -\omega_s C_0 V_m \sin(\omega_s t + \varphi_s)$$

$$+ \frac{\beta C_0 V_m}{2}(\omega_s - \omega_{pump})\sin[(\omega_{pump} - \omega_s)t$$

$$+ \varphi_{pump} - \varphi_s] - \frac{\beta C_0 V_m}{2}(\omega_{pump} + \omega_s)$$

$$\times \sin[(\omega_s + \omega_{pump})t + \varphi_s + \varphi_{pump}] \quad (12.31)$$

Equation (12.31) establishes the form of the current spectrum in a parametric capacitor. In addition to a component at the signal frequency, the spectrum contains two side frequencies, $\omega_s - \omega_{pump}$ and $\omega_s + \omega_{pump}$.

The mean power drawn by a one-port $P = (VI/2)\cos\varphi$

where φ is the phase angle between voltage and current

The mean power drawn by a parametric capacitor at the signal frequency. Circuit theory tells us that for the mean power flux between a source and its load to be non-zero, it is required that, firstly, the current and voltage under harmonic conditions be described by functions of one and the same frequency, and, secondly, the phase shift between the current and the voltage be other than 90°.

As follows from (12.31), the current flowing through a parametric capacitor always contains a reactive component at the signal frequency:

$$i_{react}(t) = -\omega_s C_0 V_m \sin(\omega_s t + \varphi_s)$$

Being in quadrature with the source voltage, this current does not dissipate power on the average.

● The choice of the pump frequency

If, however, we choose a suitable pump frequency, one more current component can be added at the signal frequency. As follows from Eqs. (12.29) and (12.30), for this to happen it will suffice to set $\omega_{pump} = 2\omega_s$. Then the current through the parametric capacitor will acquire a useful component

$$i_{useful}(t) = -\frac{\beta\omega_s C_0 V_m}{2}\sin(\omega_s t + \varphi_{pump} - \varphi_s) \quad (12.32)$$

The instantaneous power in the useful component is

$$p_{useful}(t) = v(t)\, i_{useful}(t)$$

$$= -\frac{\beta\omega_s C_0 V_m^2}{2}\cos(\omega_s t + \varphi_s)\sin(\omega_s t + \varphi_{pump} - \varphi_s)$$

$$= -\frac{\beta\omega_s C_0 V_m^2}{4}[\sin(2\omega_s t + \varphi_{pump}) - \sin(2\varphi_s - \varphi_{pump})]$$

The power averaged over a period of the signal is

$$P_{av} = \frac{1}{T_s}\int_0^{T_s} p_{useful}(t)\,dt$$

$$= \frac{\beta\omega_s C_0 V_m^2}{4}\sin(2\varphi_s - \varphi_{pump}) \quad (12.33)$$

▲ Solve Problems 5, 6 and 7

The equivalent circuit of a parametric capacitor. Equation (12.23) indicates that the average power may be positive or negative, depending on the phase relation between the input signal and the pump output. This implies that, through a suitable choice of the phase angles φ_s and φ_{pump}, a condition can be secured in which a parametrically controlled capacitor will behave as an active circuit element by supplying, rather than consuming, power at the signal frequency.

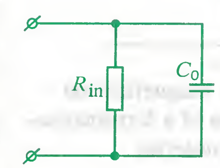
On writing $\Phi = 2\varphi_s - \varphi_{pump}$, the average power in a parametric capacitor may be written as

$$P_{av} = V_m^2/2R_{in}$$

where

$$R_{in} = 2/\beta\omega_s C_0 \sin\Phi \quad (12.34)$$

is the insertion resistance, that is, one introduced by this element into the circuit. Hence, the equivalent circuit of a parametric capacitor controlled by a pump oscillator at twice the signal frequency is a parallel connection of a capacitor C_0 and a resistor R_{in}. If this element is to act as an oscillator, it is necessary for the insertion resistance to be negative. The average power delivered into the circuit increases with decreasing magnitude of the negative resistance.



12.3 Principles of Parametric Amplification

Use of parametric amplifiers

The ability of controlled reactive one-ports to behave as active circuit elements under certain conditions is the basis of a special class of devices known as *parametric amplifiers*. They are mainly used at microwave frequencies as the input stages of highly sensitive receivers. The principal virtue of parametric amplifiers is the low level of internal noise, since they contain no sources of shot noise.

The varactor

Implementation of parametrically controlled reactive elements. Theoretically, the feasibility of parametric amplification was proved early in this century. It was not until the 1950s when the early successful parametric crystal diodes had been developed, that the idea could be realized. For their operation, parametric crystal diodes, more frequently called *varactors*, depend on the following effect. If we apply a voltage in reverse polarity to the *P-N* junction of a varactor, the separated charge q in the barrier layer will be a function of the applied voltage v. The relation $q(v)$ is called the *charge-voltage characteristic* of a varactor. A change in the applied voltage causes a displacement current to flow across the reverse-biased junction

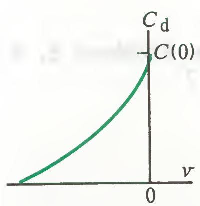
The charge-voltage characteristic

$$i = dq/dt = (dq/dv)(dv/dt) = C_d(v)\,dv/dt \tag{12.35}$$

where $C_d(v)$ is the dynamic (or incremental) capacitance of the varactor. It has been found that the dynamic (incremental) capacitance (pF) of a point-contact varactor is approximately given by

$$C_d(v) = C(0)/\sqrt{1+2|v|} \tag{12.36}$$

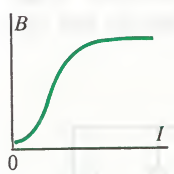
where $C(0)$ is the dynamic capacitance at zero bias voltage, and v is the applied voltage, V. The higher the bias voltage, the smaller the dynamic capacitance.



A parametrically controlled capacitance is implemented owing to the nonlinear behaviour of the varactor

Present-day varactors show a high-quality performance and can operate at frequencies up to tens of gigahertz.

It is also possible to build elements with a parametrically controlled inductance, $L(t)$. These are coils with cores made of a ferromagnetic material whose magnetic induction B strongly depends on the bias current I. At radio frequencies, however, variable-inductance elements (frequently called *varindors*) have not found any appreciable use because the cyclic magnetization of the material entails a marked time lag.



The magnetization curve of a ferromagnetic material

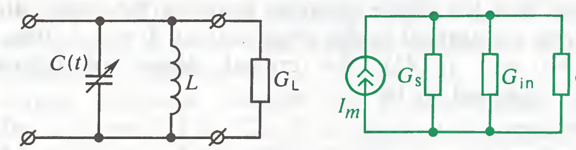
The single-stage parametric amplifier. Consider the signal source formed by a conductance G_s connected in parallel with an ideal source of current with amplitude I_m. The signal source is loaded into a resistive load whose conductance is G_L. The voltage appearing between the circuit terminals is

$$V_m = I_m/(G_s + G_L)$$

so the power in the load is

$$P_L = \frac{1}{2}\frac{I_m^2 G_L}{(G_s + G_L)^2} \tag{12.37}$$

A decrease in the conductance of the signal source leads to an increase in the power in the load. Thus, the effect of amplification can be achieved by reducing in some way the internal conductance of the signal source. One way to achieve this is to connect the signal source in parallel with a parametric capacitor whose capacitance varies at twice the signal frequency, and the initial phase of the signal is chosen such that the insertion resistance R_{in} (see Eq. (12.34)) is negative. This principle is implemented in the simple single-stage parametric amplifier shown schematically in Fig. 12.2.



(*a*) (*b*)

Fig. 12.2 Single-stage parametric amplifier: (*a*) schematic diagram; (*b*) equivalent circuit

When the stated conditions are satisfied, the parametric amplifier is said to operate in the synchronous mode

Referring to Fig. 12.2, the auxiliary inductance L and the capacitance C_0 (see Eq. (12.27)) make up a parallel resonant circuit tuned to the signal frequency. The resonant circuit has a very high input impedance, so it does not shunt the conductance inserted by the varactor in parallel with the resonant circuit and equal to

$$G_{in} = 1/R_{in} = \beta\omega_s C_0 \sin\Phi/2 \tag{12.38}$$

Referring to Fig. 12.2*b*, it is seen that the power dissipated in the load is now equal to

$$P'_L = \frac{1}{2}\frac{I_m^2 G_L}{(G_L + G_s + G_{in})^2} \tag{12.39}$$

If $G_{in} < 0$, then $P'_L > P_L$. From a comparison of Eqs. (12.37) and (12.39), the following expression can be written for the power gain factor

$$K_P = \left(\frac{G_s + G_L}{G_L + G_s + G_{in}}\right)^2 \tag{12.40}$$

As an example, for a parametric amplifier with $G_L = 0.013$ S, $G_s =$

Solve Problem 8

$= 0.01$ S, and $G_{in} = -0.02$ S, the power gain factor is

$K_P = 0.023^2/0.003^2 = 58.78$

or, on the decibel scale,

$\Delta_P = 10 \log_{10} K_P = 17.69$ dB

The stability of a parametric amplifier. Should the negative conductance of the varactor completely balance both the internal conductance of the signal source and the load conductance, the parametric amplifier will jump into self-excited oscillations. Equation (12.39) shows that this occurs at the critical value of the negative insertion conductance, defined by

$$G_{in,cr} = -(G_s + G_L) \tag{12.41}$$

The condition for the self-excitation of the parametric amplifier

Assuming that the phase relations between the signal and the pump output are optimal in the sense that $\sin \Phi = -1$, then, from Eqs. (12.34) and (12.41), the critical degree of capacitance modulation is found to be

$$\beta_{cr} = \frac{2(G_s + G_L)}{\omega_s C_0} \tag{12.42}$$

Example 12.4. *A single-stage parametric amplifier operates at 6 GHz ($\lambda = 5$ cm). The signal source and the load have the same conductance of 0.005 S ($R_s = R_L = 200\ \Omega$). The capacitance of the varactor is $C_0 = 0.8$ pF. Find the limits of capacitance variation at which the circuit jumps into self-excited oscillation.*

From Eq. (12.42), we get

$$\beta_{cr} = \frac{2 \times 0.01}{6.28 \times 6 \times 10^9 \times 0.8 \times 10^{-12}} = 0.66$$

Thus, the parametric amplifier will jump into self-excited oscillations if the capacitance of the varactor varies harmonically in time between the limits $C_{max} = C_0(1 + \beta_{cr}) = 1.328$ pF and $C_{min} = C_0(1 - \beta_{cr}) = 0.272$ pF.

Parametric amplification in the asynchronous condition. In practice, it is difficult, if at all possible, to satisfy the condition of synchronism, $\omega_{pump} = 2\omega_s$, exactly. If the signal frequency is somewhat off the desired value, that is, if

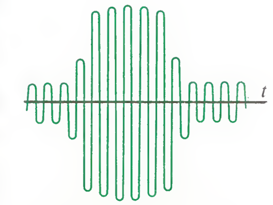
$$\omega_s = \omega_{pump}/2 + \delta\omega$$

the parametric amplifier is said to operate in the *asynchronous mode*. Now Φ, which, in accord with Eq. (12.34), controls the value of the insertion resistance, is a function of time:

$$\Phi = 2\omega_s - \omega_{pump} - \delta\omega t$$

Varying as

$$R_{in} = \frac{2}{\beta\omega_s C_0 \sin(2\omega_s - \omega_{pump} - \delta\omega t)}$$



The output signal of a parametric amplifier in the asynchronous condition

the insertion resistance changes sign periodically. In consequence, the output signal undergoes deep variations in level, similar to beats. This drawback of single-stage parametric amplifiers is a serious obstacle to their practical use.

The double-stage parametric amplifier. Attempts to improve the performance of parametric amplifiers have resulted in fundamentally distinct parametric amplifier designs which are free from the above shortcomings. These are *double-stage parametric amplifiers* (up-converters and down-converters) which can operate with any phase relation between the signal and pump frequencies and irrespective of the initial phases of the two waves. This effect is achieved through the use of an auxiliary oscillation at one of the intermodulation frequencies.

The idling (idler) circuit and the idler frequency

A circuit schematic diagram of a double-stage parametric amplifier is shown in Fig. 12.3. It is seen that the amplifier consists of two tuned circuits. One, called the *signal tuned circuit*, is tuned to ω_s; the other, called the *idling* (or *idler*) *circuit*, is tuned to the *idler frequency* $\omega_{idler} = \omega_{pump} - \omega_s$. The two circuits are coupled by a varactor whose capacitance varies in time harmonically at the pump frequency ω_{pump}:

$$C(t) = C_0(1 + \beta \cos \omega_{pump} t)$$

If $v_s(t)$ and $v_{idler}(t)$ are voltages across the signal circuit and the idler circuit, respectively, then the current through the varactor is

It is assumed that the modulation depth of the capacitance is sufficiently small

$$i(t) \approx C(t) \frac{d}{dt}(v_s - v_{idler}) \tag{12.43}$$

Suppose that

$$v_s(t) = V_{m,s} \cos \omega_s t \tag{12.44}$$

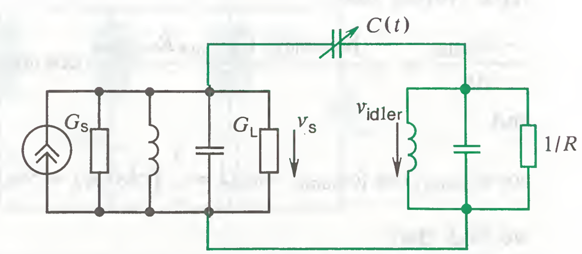


Fig. 12.3 Double-stage parametric amplifier

The intermodulation components of the current flowing through the varactor include a component at the idler frequency

$$i_{\text{idler}}(t) = I_{m,\text{idler}} \cos(\omega_{\text{idler}} t + \Psi) \tag{12.45}$$

where $I_{m,\text{idler}}$ and Ψ remain to be defined yet.

Assuming that we know $R_{\text{res,idler}}$, the resonant resistance of the idler circuit, we may write

$$v_{\text{idler}}(t) = I_{m,\text{idler}} R_{\text{res,idler}} \cos(\omega_{\text{idler}} t + \Psi) \tag{12.46}$$

On substituting (12.44) and (12.46) in (12.43), we obtain the following expression for the current in the parametric capacitor:

$$i(t) = C_0(1 + \beta\cos\omega_{\text{pump}} t)[-\omega_s V_{m,s} \sin\omega_s t + \omega_{\text{idler}} I_{m,\text{idler}} R_{\text{res,idler}} \sin(\omega_{\text{idler}} t + \Psi)]$$

In order to derive $i_{\text{idler}}(t)$ from the above expression, we note that the component at $\omega_{\text{idler}} = \omega_{\text{pump}} - \omega_s$ can be obtained only from the product

$$\cos\omega_{\text{pump}} t \sin\omega_s t = {}^1/_2 [\sin(\omega_s - \omega_{\text{pump}}) t + \sin(\omega_s + \omega_{\text{pump}}) t]$$

Thus, the current at the idler frequency does not depend on the voltage across the idler circuit, $v_{\text{idler}}(t)$, and is equal to

The system acts as a current source

$$i_{\text{idler}}(t) = \frac{\beta\omega_s C_0 V_{m,s}}{2} \sin\omega_{\text{idler}} t \tag{12.47}$$

Therefore,

$$v_{\text{idler}}(t) = \frac{\beta\omega_s C_0 V_{m,s} R_{\text{res,idler}}}{2} \sin\omega_{\text{idler}} t$$

Now let us find the current through the varactor at the signal frequency, $i_s(t)$. By analogy with the previous reasoning, we find that this current is independent of the voltage across the signal circuit, $v_s(t)$. Noting that

$$\frac{-dv_{\text{idler}}}{dt} = -\frac{\beta\omega_s \omega_{\text{idler}} C_0 V_{m,s} R_{\text{res,idler}}}{2} \cos\omega_{\text{idler}} t$$

and

$$\cos\omega_{\text{pump}} t \cos(\omega_{\text{pump}} - \omega_s) t = \frac{1}{2}[\cos\omega_s t + \cos(2\omega_{\text{pump}} - \omega_s) t]$$

we find that

$$i_s(t) = \frac{-\beta^2\omega_s \omega_{\text{idler}} C_0^2 V_{m,s} R_{\text{res,idler}}}{4} \cos\omega_s t \tag{12.48}$$

Therefore, the conductance introduced in the signal circuit by the series combination of a varactor and the idler circuit is equal to

$$G_{\text{in}} = \frac{i_s(t)}{v_s(t)} = -\frac{\beta^2\omega_s \omega_{\text{idler}} C_0^2 R_{\text{res,idler}}}{4} \tag{12.49}$$

Since the insertion conductance is negative, the circuit is capable of amplifying the signal in power. The gain factor is found by Eq. (12.40). Stability analysis is carried out in the same way as for the single-stage parametric amplifier.

Advantages of a double-stage parametric amplifier

If we compare Eqs. (12.38) and (12.49), we will above all see that in a double-stage parametric amplifier the negative insertion conductance is in no way related to the initial phases of the signal and the pump output. Also, the precise choice of ω_s and ω_{pump} is not critical. The insertion conductance will always be negative so long as $\omega_{\text{pump}} > \omega_s$.

Power balance in multistage parametric amplifiers. The insensitivity of a double-stage parametric amplifier towards the phase relation between the signal and the pump output makes it possible to analyse such systems on the basis of simple power relations. Let us turn to the general circuit diagram in Fig. 12.4.

The reasoning equally applies if a system contains a nonlinear inductive element

Here, three circuits are connected in parallel with a nonlinear capacitance C_{nl}. One is the signal circuit, the second is the pump circuit, and the third is the idler circuit tuned to an intermodulation frequency

$$\omega_i = m\omega_s + n\omega_{\text{pump}} \quad (m \text{ and } n \text{ are integers})$$

Each circuit includes a narrowband filter which transmits only frequencies close to ω_s, ω_{pump}, and ω_i, respectively. For simplicity, let the signal and pump circuits be free from ohmic losses.

Let us omit for a moment the signal source or the pump. Then the current flowing through the nonlinear capacitor will not carry components at intermodulation frequencies. The current in the idler

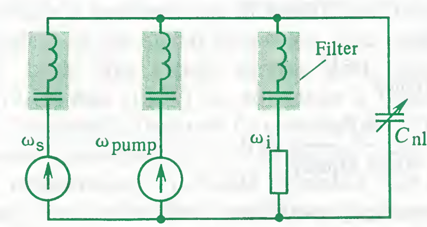


Fig. 12.4 To the derivation of power relations for a two-stage parametric system

circuit will be zero, and the system as a whole will behave as a reactive network which draws no average power from the source.

Now we let both the signal source and the pump be present in the circuit. In the circumstances, the current through the nonlinear capacitor includes a component at an intermodulation frequency; this current can complete its path only through the idler circuit. The load in the idler circuit draws an average power, and positive or negative resistances are inserted in the signal and pump circuits, the magnitude and sign of which govern the distribution of power between the signal source and the pump.

This is a closed (self-contained) system and, on the basis of the law of conservation of energy, the average powers in the signal, the pump output, and the intermodulation frequency are connected by a relation of the form

$$P_s + P_{pump} + P_i = 0 \tag{12.50}$$

The power averaged over period T may be expressed in terms of the energy E dissipated over this time interval as

$$P = \frac{1}{T}\int_0^T v(t)\,i(t)\,dt = fE$$

where f is the frequency in Hz. Thus,

$$f_s E_s + f_{pump} E_{pump} + f_i E_i = 0$$

or, recalling that $f_i = mf_s + nf_{pump}$,

$$f_s(E_s + mE_i) + f_{pump}(E_{pump} + nE_i) = 0 \tag{12.51}$$

Equality (12.51) must be satisfied identically at any values of f_s and f_{pump}. This can be so only if

$$\left.\begin{aligned} &E_s + mE_i = 0 \\ &E_{pump} + nE_i = 0 \end{aligned}\right. \tag{12.52}$$

By replacing energies with powers, we obtain two important equations known as the *Manley-Rowe relations*:

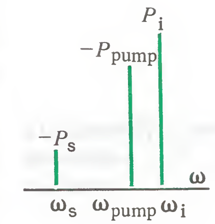
The Manley-Rowe relations

$$\left.\begin{aligned} &P_s/f_s + \frac{mP_i}{mf_s + nf_{pump}} = 0 \\ &P_{pump}/f_{pump} + \frac{nP_i}{mf_s + nf_{pump}} = 0 \end{aligned}\right. \tag{12.53}$$

The Manley-Rowe relations give a simple and easy-to-grasp insight into power transformation by multistage parametric systems. Let us illustrate this with two typical cases.

Parametric up-conversion. On setting $m = n = 1$ in Eqs. (12.53), we obtain

$$\left.\begin{aligned} &P_s/f_s + P_i/(f_s + f_{pump}) = 0 \\ &P_{pump}/f_{pump} + P_i/(f_s + f_{pump}) = 0 \end{aligned}\right. \tag{12.54}$$



As is customary, let the power dissipated in the load be positive, and that supplied by the signal source, negative. From Eq. (12.54) it is seen that since $P_i > 0$, then $P_s < 0$ and $P_{pump} < 0$. Thus, if the idler circuit of the amplifier is tuned to a frequency $f_i = f_s + f_{pump}$, both the signal source and the pump will deliver power to the idler circuit where it is used up in the load. Since

$$P_i = -P_s - P_{pump}$$

the power gain factor is

$$K_P = P_i/-P_s = 1 + f_{pump}/f_s = f_i/f_s \tag{12.55}$$

An advantage of the parametric up-converter is the absolute stability of the system which will not jump into self-excited oscillation at any value of signal and pump power. A disadvantage of the circuit is that the output signal has a higher frequency than the input signal. At microwave frequencies, this brings about certain difficulties during the subsequent signal processing.

Regenerative parametric amplification. Let $m = -1$ and $n = 1$, which means that the idler (output) circuit is tuned to $f_i = f_{pump} - f_s$. The Manley-Rowe relations now take the form

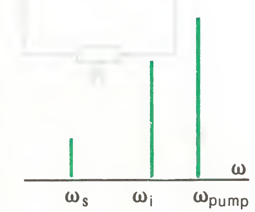
$$P_s/f_s - P_i/(f_{pump} - f_s) = 0 \tag{12.56}$$

and

$$P_{pump}/f_{pump} + P_i/(f_{pump} - f_s) = 0 \tag{12.57}$$

It is assumed that $f_{pump} > f_s$

As follows from Eq. (12.56) now both P_i and P_s are positive. Thus, some of the power drawn from the pump is transferred to the signal circuit which is another way of saying that regeneration takes place in the system at the signal frequency. The output power may be taken off both the signal circuit and the idler circuit.



Equations (12.56) and (12.57) do not make it possible to find the gain factor of the system, because P_s contains both the part drawn from the devices connected to the amplifier input, and the part arising due to regeneration. It should be noted that such amplifiers are apt to jump into self-excited oscillation, because, given certain conditions, a non-zero power will be developed in the signal circuit even in the absence of the signal applied to the amplifier input from external circuits.

12.4 Nonstationary Dynamic Systems

Linear dynamic systems containing parametric elements are a more complicated form of linear parametric devices. Their models may be *RC*-, *RL*- or *RCL*-networks with circuit elements of the form $R(t)$, $C(t)$ or $L(t)$. In the general case, such systems can be described mathematically, using linear differential equations with variable coefficients:

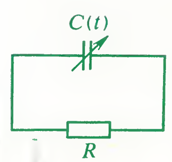
$$a_n(t)\frac{d^n v_{out}}{dt^n} + a_{n-1}(t)\frac{d^{n-1} v_{out}}{dt^{n-1}} + \dots$$
$$+ a_1(t)\frac{dv_{out}}{dt} + a_0(t)v_{out} = f(t) \qquad (12.58)$$

General methods for an analytic solution of such equations, given an arbitrary dependence of the coefficients on time, are nonexistent. This is why a prominent role is played by approximate methods based on the specific behaviour of a particular system. This section will examine two nonstationary dynamic systems for one of which a rigorous solution can be obtained, whereas the other permits us to find with a relative ease the approximate characteristics of the output signal of importance to practical applications.

Free oscillations in a parametric RC-network. Consider a source-free *RC*-network made up of a fixed resistor R and a parametric capacitor of capacitance $C(t)$. Initially, the voltage across the capacitor is $v_C(0) = V_0$. We set out to find the manner in which v_C varies for $t > 0$. To write a differential equation for the system, we note that the current in the circuit is

$$i(t) = C\,dv_C/dt + v_C\,dC/dt$$

By Kirchhoff's second law, the initial-value problem may be stated as



$$\begin{cases} RC\,dv_C/dt + (1 + R\,dC/dt)v_C = 0 \\ v_C(0) = V_0 \end{cases} \qquad (12.59)$$

Importantly, if the capacitance were fixed, the initial-value problem for the system in question would be stated as

$$\begin{cases} RC\,dv_C/dt + v_C = 0 \\ v_C(0) = V_0 \end{cases} \qquad (12.60)$$

By comparing (12.59) and (12.60), we note that in the general case the differential equation of a parametric system cannot be developed by simply inserting the variables $R(t)$, $C(t)$ or $L(t)$ in the equation corresponding to a stationary system.

Consider the specific case of a linearly varying capacitance, when

$$C(t) = C_0 + at$$
$$dC/dt = a$$

where C_0 and a are certain constants.

The differential equation can readily be solved by separating the variables:

The quantity a has the dimensions of conductance

$$\frac{dv_C}{v_C} = d\ln v_C = -\frac{1 + aR}{R(C_0 + at)}dt$$

Hence,

$$\ln v_C = -\frac{1 + aR}{R}\int\frac{dt}{C_0 + at} = -\frac{1 + aR}{R}\ln(at + C_0) + \ln A$$

or

$$v_C(t) = A(at + C_0)^{-(1+aR)/aR} \qquad (12.61)$$

The constant A is readily found from the initial condition

$$v_C(0) = AC_0^{-(1+aR)/aR} = V_0$$

The final solution of the problem (12.59) has the form

$$v_C(t) = V_0(1 + at/C_0)^{-(1+aR)/aR} \qquad (12.62)$$

It is useful to compare Eq. (12.62) with the solution for a stationary problem for $d = 0$:

$$v_C(t) = V_0\exp(-t/RC_0) \qquad (12.63)$$

The plots constructed on the basis of Eqs. (12.62) and (12.63) appear in Fig. 12.5. The parameter a has been taken as $1/R$. In the circumstances, the capacitance of the capacitor increases by C_0 over

▲ **Work Problem 9**

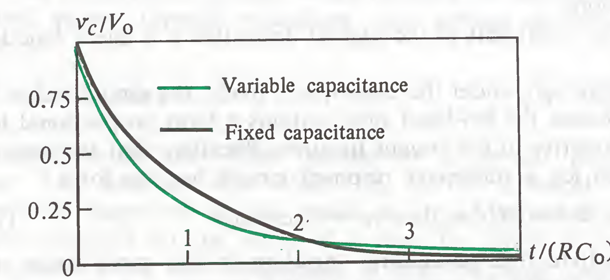
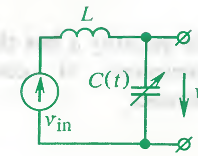


Fig. 12.5 Variations in the voltage across the capacitor of a parametric *RC*-network during a source-free discharge. It also shows the curve related to a stationary circuit with a time constant $\tau_0 = RC_0$

the time interval $\tau_0 = RC_0$ which is the time constant of a stationary circuit.

Interestingly, the voltage across the parametric system varies at a higher rate during the initial time interval, but later the exponential (12.63) falls off more rapidly than the power function (12.62).



A tunable-capacitance resonant circuit. Quite a number of communication devices, such as spectrum analyzers or scanning receivers, make use of resonant circuits made up of an L and a C, one of which is made to vary in time by some means, thereby varying the resonant frequency of the tuned circuit.

Consider a tunable resonant circuit under the following conditions: (1) The parametric element is a capacitor from which the output signal is picked off. (2) There are no losses in the resonant circuit. (3) The input signal is supplied by a source of harmonic voltage $v_{in}(t) = V_m \cos \omega_0 t$.

Recalling that the voltage across an inductor is given by

$$L\,di/dt = L\frac{d}{dt}(C\,dv_C/dt + v_C\,dC/dt)$$

$$= LC\,d^2v_C/dt^2 + 2L(dC/dt)(dv_C/dt) + L\,v_C(d^2C/dt^2)$$

we derive the following differential equation of the system:

$$LC\,d^2v_C/dt^2 + 2L(dC/dt)(dv_C/dt) + (L\,d^2C/dt^2 + 1)\,v_C = v_{in}(t) \tag{12.64}$$

If the capacitance varies linearly in time as $C(t) = at$, then Eq. (12.64) can be simplified as

$$aLT\,d^2v_C/dt^2 + 2aL\,dv_C/dt + v_C = V_m \cos \omega_0 t \tag{12.65}$$

The linear differential equation thus obtained has two distinctions.

(1) The coefficient of the highest derivative is a linear function of time.

(2) Although, under the assumption made, the circuit is free from ohmic losses, the left-hand side contains a term proportional to the first derivative of the sought function. Recalling that the analogous equation for a stationary resonant circuit has the form

$$LC\,d^2v_C/dt^2 + RC\,dv_C/dt + v_C = V_m \cos \omega_0 t \tag{12.66}$$

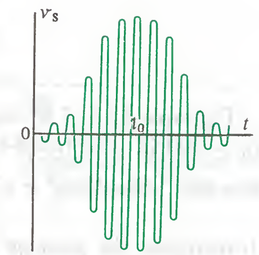
we conclude that parametric variation in the capacitance of the resonant circuit is equivalent to the insertion of an amount of attenuation.

Generally speaking, Eq. (12.65) must be supplemented with initial conditions defining v_C and dv_C/dt for $t = 0$. The total solution (see

Chap. 8) is the sum of a particular integral and the complementary function. The particular integral represents the forced (steady-state) response of the system, and the complementary function its free (transient) response. We will limit ourselves to seeking the forced response as being of the greatest practical interest.

Suppose that in the sense to be defined shortly the capacitance $C(t)$ varies slowly. This permits us to speak of the *instantaneous resonant frequency*

$$\omega_{res}(t) = 1/\sqrt{LC(t)} = 1/\sqrt{Lat} \tag{12.67}$$



At time t_0, when $\omega_{res}(t_0) = \omega_0$, the system is at resonance, and the amplitude of the output wave is a maximum. At $t < t_0$ and $t > > t_0$, the output signal has a lower amplitude.

Since we are interested in the behaviour of the system in the vicinity of the resonant point t_0, let us replace approximately the coefficient aLt on the left-hand side of Eq. (12.65) with a constant

$$aLt_0 = 1/\omega_0^2 \tag{12.68}$$

Now the problem reduces to finding the partial solution of the differential equation

$$d^2v_C/dt^2 + 2aL\omega_0^2\,dv_C/dt + \omega_0^2\,v_C = \omega_0^2\,V_m \cos \omega_0 t \tag{12.69}$$

Since the capacitance is assumed to vary slowly, the problem reduces to solving a differential equation with constant coefficients

We will seek the solution in the form

$$v_C(t) = A\cos \omega_0 t + B \sin \omega_0 t \tag{12.70}$$

with unknown constants A and B.

On substituting (12.70) in (12.69) and equating the coefficients of the sine and the cosine on both sides, we get

$$A = 0, \quad B = V_m/2aL\omega_0$$

Thus, over a short time interval which includes the instant when the system passes through resonance, the voltage across the capacitor is

$$v_C(t) \approx (V_m/2aL\omega_0) \sin \omega_0 t \tag{12.71}$$

The condition for the "slow" passage of a system through resonance. The accuracy with which Eq. (12.71) represents the actual process improves as the value of a is decreased. The constraints imposed on the value of a may be stated in more rigorous terms if we take advantage of the following physical consideration: The passage of a parametric system through resonance should be regarded as slow, if over the period $T = 2\pi/\omega_0$ the instantaneous resonant frequency $\omega_{res}(t) = 1/\sqrt{LC(t)}$ changes by a substantially smaller amount than ω_0.

In many cases of practical interest the condition of slow passage is satisfied

The change in the resonant frequency is

$$\Delta\omega_{res} = \Delta\omega_{res}/dt\,|_{t=t_0} \cdot T = -aL\omega_0^3 T/2 = -a\pi L\omega_0^2 \qquad (12.72)$$

The condition will be satisfied if

$$a\pi L\omega_0^2 \ll \omega_0$$

Hence,

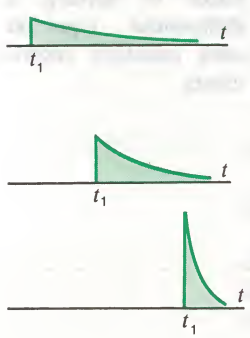
$$a \ll 1/\pi\omega_0 L = 1/\pi\rho_0 \qquad (12.73)$$

where ρ_0 is the characteristic impedance of the resonant circuit tuned to resonate at frequency ω_0.

Example 12.5. *The characteristic impedance of a resonant circuit is $\rho_0 = 500\ \Omega$. The tuned-circuit capacitor changes in value at the rate $a = 10^{-4}$ S $= 10^{-4}$ F s^{-1} = 100 pF μs^{-1}*

Investigate the passage of the system through resonance.

Taking advantage of the inequality (12.73), we see that the condition for the slow change of the capacitance is satisfied. On the basis of (12.71), we find that at resonance the amplitude of the output voltage is ten times the amplitude of the input signal $1/2aL\omega_0 = 1/2a\rho_0 = 10$



Variations in the shape of the impulse response of a parametric network with time

The impulse response and the frequency response of a nonstationary dynamic system. Any linear system, be it stationary or nonstationary (parametric), obeys the principle of superposition, and so it may be investigated by recourse to the Duhamel superposition integral.

The impulse response of a parametric system, $h(t, t_1)$ is defined as the response to the delta-impulse applied to the input at time t_1. Whereas for a stationary system the argument of the impulse response is solely the difference, $t - t_1$, between the instant of observation and the instant when the excitation is applied, for a nonstationary system this relation may take an arbitrary form. Still, a physical system cannot give an output in advance of the input signal, so for a physically realizable system we always have that

$$h(t, t_1) = 0 \text{ for } t < t_1 \qquad (12.74)$$

If we know the impulse response $h(t, t_1)$, we can readily define the output signal of a nonstationary dynamic system as

$$v_{out}(t) = \int_{-\infty}^{\infty} v_{in}(t_1) h(t, t_1)\,dt_1 \qquad (12.75)$$

The Fourier transform of the impulse response in terms of the variable t_1 is called the *frequency response of a parametric network*:

$$K(j\omega, t) = \int_{-\infty}^{\infty} h(t, t_1)\exp(-j\omega t_1)\,dt_1 \qquad (12.76)$$

If $S_{in}(\omega)$ is the spectrum of the input signal, then Eq. (12.75) in frequency representation takes the form

$$v_{out}(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} S_{in}(\omega) K(j\omega, t)\exp(j\omega t)\,d\omega \qquad (12.77)$$

A distinction of the frequency response of a parametric network

A distinction of the frequency response (12.76) from the similar function characterizing a stationary system consists in that it has an additional argument – the observation time t.

Actually, Eq. (12.75) or (12.77) is a formal solution because the evaluation of the impulse response $h(t, t_1)$ or, which is the same, of the frequency response $K(j\omega, t)$ involves an analysis of an exact model of the system, that is, solving the corresponding differential equation. Yet, this device is frequently employed when a complex real network has to be replaced with a maximally simplified mathematical model. For example, a parametric system may have the frequency response

$$K(j\omega, t) = z(t) \qquad (12.78)$$

which is independent of frequency. The corresponding device operates as a multiplier or an amplitude modulator. When it is driven by an input wave of the form

$$v_{in}(t) = V_0 \cos\omega_0 t$$

its output signal is an AM wave

$$v_{out}(t) = V_0 z(t)\cos\omega_0 t \qquad (12.79)$$

The simplifications assumed in describing nonstationary dynamic systems

Another important example is a nonstationary system operating as a frequency or phase modulator. Here

$$K(j\omega, t) = \exp[jz(t)] \qquad (12.80)$$

and so, under the conditions defined above, the output signal

$$v_{out}(t) = V_0 \cos[\omega_0 t + z(t)] \qquad (12.81)$$

is an angle-modulated wave.

12.5 Response of Parametric Systems with Random Characteristics to Harmonic Signals

It is of great interest, both theoretically and in the applied sense, to investigate the response of a system whose parameters vary in time at random. In the simplest case, this may be a random instability in the frequency response function of a device, leading to fluctuations in the amplitude at the output. In a more complicated situation, one may be concerned with the propagation of signals in various media, say, in the Earth's ionosphere subject to variations in the refractive index. Here the received signal is corrupted by a random angle modulation because the shift of the signal phase along the propagation path is a random time function.

In most real cases the signal is both amplitude and phase modulated at random

A detailed investigation of the statistical characteristics of the signals at the output of linear systems with randomly varying parameters is an extremely complex task [14, 15]. In this section, we will only dwell in brief on two of the simplest problems of the kind.

Random amplitude modulation. In investigating a parametric system whose frequency response has the form defined in (12.78), we assume to know the mean m_z and the autocorrelation function $K_z(\tau)$ of the stationary random process $Z(t)$ which defines the form of modulation.

By resort to correlation theory, we will investigate the statistical characteristics of the process $Y(t)$ at the output of a system, assuming that the input signal is a harmonic wave, $V_0 \cos \omega_0 t$. Since the realization of the output signal is

$$y(t) = V_0 z(t) \cos \omega_0 t$$

it follows that $m_y = \bar{y} = 0$.

The autocorrelation function of the output signal is

$$K_y(\tau) = \overline{y(t)\, y(t+\tau)} = \overline{V_0^2 z(t)\, z(t+\tau) \cos \omega_0 t \cdot \cos \omega_0 (t+\tau)} \tag{12.82}$$

Since

$$\overline{\cos \omega_0 t \cos \omega_0 (t+\tau)} = 1/2 \overline{[\cos \omega_0 (2t+\tau) + \cos \omega_0 \tau]} = 1/2 \cos \omega_0 \tau$$

then

$$K_y(\tau) = (V_0^2/2)\, \overline{z(t)\, z(t+\tau)} \cos \omega_0 \tau$$

By definition, the mean product is

$$\overline{z(t)\, z(t+\tau)} = K_z(\tau) + m_z^2$$

Hence, the final relation connecting the autocorrelation function of the output signal and the random frequency response $z(t)$ takes the form

$$K_y(\tau) = (V_0^2/2)\,[K_z(\tau) + m_z^2] \cos \omega_0 \tau \tag{12.83}$$

and the variance of the output random process is

$$\sigma_y^2 = (V_0^2/2)(\sigma_z^2 + m_z^2)$$

The form of Eq. (12.83) is an indication that if the realizations $z(t)$ vary more slowly than the input signal, the output wave is a narrowband random process. It is to be noted that if $m_z \neq 0$, then $K_y(\tau)$ does not tend to zero for $\tau \to \infty$.

To get insight into the physical significance of the above property, let us take the inverse Fourier transform of $K_y(\tau)$, that is, find the power spectrum of the process $Y(t)$:

$$W_y(\omega) = \int_{-\infty}^{\infty} K_y(\tau) \exp(-j\omega\tau)\, d\tau$$

$$= \frac{V_0^2 \sigma_z^2}{2} \int_{-\infty}^{\infty} R_z(\tau) \cos \omega_0 \tau \cos \omega \tau \, d\tau$$

$$+ \frac{V_0^2 m_z^2}{2} \int_{-\infty}^{\infty} \cos \omega_0 \tau \, d\tau$$

By simple rearrangement, we can re-write the above equation as

$$W_y(\omega) = \frac{V_0^2 \sigma_z^2}{2} \Big[\int_{-\infty}^{\infty} R_z(\tau) \cos(\omega_0 + \omega)\, \tau d\tau$$

$$+ \int_{-\infty}^{\infty} R_z(\tau) \cos(\omega_0 - \omega)\, \tau d\tau \Big]$$

$$+ \frac{V_0^2 m_z^2 \pi}{2} [\delta(\omega + \omega_0) + \delta(\omega - \omega_0)] \tag{12.84}$$

Equation (12.84) tells us that the power spectrum of the process at the output of a random amplitude modulator contains two components: a continuous part due to random fluctuations in the amplitude, and a discrete part which represents the response of the system to the unmodulated carrier wave; the spectrum of the discrete part corresponds to two delta-functions in the frequency domain. The contribution of the discrete part increases as m_z^2 becomes progressively greater than σ_z^2. In approximate form, the autocorrelation function and the power spectrum for the case in question are plotted in Fig. 12.6. As is seen, the continuous part of the spectrum is represented by two smooth curves with maxima at points $\omega = \pm \omega_0$.

The power spectrum of a signal in the case of random amplitude modulation

The envelope of the output signal. If $Z(t)$ is a slow process, it is

The envelope

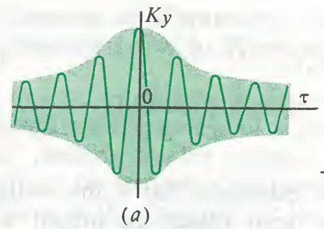
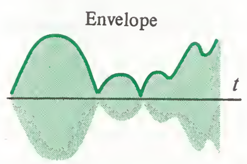
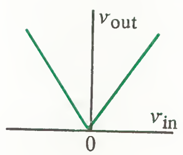


Fig. 12.6 Characteristics of the signal at the output of a random amplitude modulator: (a) autocorrelation function; (b) power spectrum





legitimate to think that the instantaneous value of the physical envelope at the output of the system is

$$V_y(t) = |V_0 z(t)| \tag{12.85}$$

The modulus sign indicates that the amplitude detector producing this envelope at its output is insensitive to the phase of the high-frequency carrier.

As is seen from (12.85), the envelope of the narrowband process at the output of a parametric system with a fluctuating frequency response $z(t)$ may be obtained as a result of a nonlinear lag-free (zero-memory) transformation of the random process $z(t)$ in an imaginary device with a piecewise-linear characteristic $v_{out} = |V_0 v_{in}|$. The mean, the variance, and the autocorrelation function of the envelope can be calculated by the techniques set forth in Chap. 11.

Random angle modulation. Now let us turn to a random parametric system whose frequency response has the form defined in (12.80). Going through the steps taken in our investigation of a random amplitude modulator, we can develop a general expression for the autocorrelation function of the output signal when the system is driven by a harmonic excitation:

$$K_y(\tau) = V_0^2 \overline{\cos[\omega_0 t + z(t)] \cos[\omega_0(t+\tau) + z(t+\tau)]}$$

$$= (V_0^2/2)\{\overline{\cos[2\omega_0 t + \omega_0\tau + z(t) + z(t+\tau)]}$$

$$+ \overline{\cos[\omega_0\tau + z(t+\tau) - z(t)]}\} \tag{12.86}$$

The variance of the signal remains constant in the case of random angle modulation

On averaging, the first term in the braces will obviously vanish, and so

$$K_y(\tau) = (V_0^2/2)\overline{\cos(\omega_0\tau + z_\tau - z)}$$

$$= (V_0^2/2)\overline{\cos(z_\tau - z)}\cos\omega_0\tau - (V_0^2/2)\overline{\sin(z_\tau - z)}\sin\omega_0\tau \tag{12.87}$$

(To simplify the expression, we have omitted the argument of the

function $z(t)$.)

If $\tau \to 0$, then

$$\lim \overline{\cos(z_\tau - z)} = 1$$

and

$$\lim \overline{\sin(z_\tau - z)} = 0$$

so that the effective power in the signal, that is, its variance $\sigma_y^2 = V_0^2/2$ is the same as the power in the harmonic signal of amplitude V_0 and at a constant frequency.

Equation (12.87) gives a complete description of a signal subjected to random angle modulation in terms of correlation theory. Among other things, it tells us that if the process $Z(t)$ is formed by realizations which are slow in comparison with harmonic waves at frequency ω_0, then the output signal of a random phase modulator is a narrowband process with a centre frequency ω_0.

Angle modulation by a normal random process. In order to make use of Eq. (12.87), we should find the mean of its trigonometric functions with a difference argument. This can be done on the basis of a bivariate probability density function $p(z_\tau, z)$:

$$\overline{\cos(z_\tau - z)} = \iint_{-\infty}^{\infty} \cos(z_\tau - z)\, p(z_\tau, z)\, dz_\tau dz$$

$$\overline{\sin(z_\tau - z)} = \iint_{-\infty}^{\infty} \sin(z_\tau - z)\, p(z_\tau, z)\, dz_\tau dz$$

It may prove difficult to evaluate such integrals if the function $p(z_\tau, z)$ is arbitrary. If, however, $Z(t)$ is a Gaussian process, an elegant technique suggests itself, immediately leading to the final result. It is based on the use of a bivariate characteristic function for a Gaussian process (see Eq. (6.32)):

Averaging with the aid of the characteristic function

$$\Theta(v_1, v_2) = \overline{\exp[j(z_\tau v_1 + z v_2)]}$$

$$= \exp\left[jm_z(v_1 + v_2) - \frac{1}{2}\sigma_z^2(v_1^2 + 2R_z(\tau)v_1 v_2 + v_2^2)\right] \tag{12.88}$$

Since

$$\cos(z_\tau - z) = \frac{\exp[j(z_\tau - z)] + \exp[-j(z_\tau - z)]}{2}$$

then, by virtue of Eq. (12.88),

$$\overline{\cos(z_\tau - z)} = \frac{1}{2}[\Theta(1, -1) + \Theta(-1, 1)] \tag{12.89}$$

On putting for definiteness $m_z = 0$, we have

$$\Theta(1, \ -1) = \Theta(-1, \ 1) = \exp\{-\sigma_z^2[1 - R_z(\tau)]\}$$

Therefore,

$$\overline{\cos(z_\tau - z)} = \exp\{-\sigma_z^2[1 - R_z(\tau)]\}$$

$$\overline{\sin(z_\tau - z)} = 0$$

Substituting the above results in (12.87) yields the final expression for the autocorrelation function of the signal produced from a harmonic wave by Gaussian angle modulation:

$$K_y(\tau) = \frac{V_0^2}{2}\exp\{-\sigma_z^2[1 - R_z(\tau)]\}\cos\omega_0\tau \tag{12.90}$$

Qualitatively, the above function is analogous to that found earlier in the analysis of random amplitude modulation. Therefore, the conclusion that the power spectrum contains two components, continuous and discrete, may fully be carried over to the present case. Analysis [15] shows that for $\sigma_z^2 \gg 1$ random angular modulation is broadband. The discrete part of the spectrum practically disappears, whereas the continuous part in the vicinity of frequency ω_0 is described by a Gaussian function of the form

$$W_y(\omega) = V_0^2\sqrt{\pi/2}\,\frac{1}{\sigma_z\sqrt{-R_z''(0)}}\exp\frac{-(\omega - \omega_0)^2}{2\sigma_z^2[-R_z''(0)]} \tag{12.91}$$

The effective bandwidth

$$\Delta\omega_{\text{eff}} = \sqrt{2\pi}\,\sigma_z\sqrt{-R_z''(0)} \tag{12.92}$$

increases with an increase in both σ_z and $-R_z''(0)$ proportional to the time rate of change of the modulating function.

Summary

- If the impedance of a lag-free (zero-memory) parametric element periodically varies in time, the spectrum of the output signal contains, generally speaking, an infinite number of intermodulation products at sum and difference frequencies of the form $\omega_s + k\omega_c$ $(k = 0, \ \pm 1, \ \pm 2, \ \ldots)$.
- A resistive parametric element can be implemented by applying the sum of a small message signal and a large control signal to the input of a lag-free nonlinear one-port.
- Frequency conversion consists in translating the signal spectrum from the neighbourhood of the carrier frequency into the neighbourhood of the intermediate frequency without any change in the form of modulation.

- In synchronous detection, the signal frequency is the same as the local oscillator frequency.
- Reactive parametric elements can transfer some of the pump power to the circuits containing the useful signal.
- Given appropriate phase relations, a parametrically controlled capacitor can initiate oscillations in an LC resonant circuit. Connection of such a capacitor is equivalent to the insertion of a negative conductance in the resonant circuit.
- Parametric amplifiers may be single-stage and double-stage. In the latter case, a parametric amplifier includes an idler circuit tuned to one of the intermodulation frequencies.
- Power relations in a multistage parametric system are described by the Manley-Rowe relations.
- Parametric dynamic systems are described by differential equations with variable coefficients.
- The frequency response of a linear parametric system is a function of both frequency and time.
- In the case of random amplitude modulation the spectrum of the output signal contains both a continuous and a discrete component.
- If, in the case of angle modulation, the modulating function is a realization of a normal random process, the autocorrelation function of the output signal may be stated in terms of the characteristic function of the signal at the modulator input.

Review Questions

1. What is the fundamental difference between the spectra of the currents flowing in a resistive parametric one-port and a nonlinear one-port? Assume that both elements are driven by a harmonic excitation.
2. Draw up a block diagram of a superhet receiver. What is the image frequency in reception? How can the ambiguity in receiver tuning be resolved?
3. Define the conversion transconductance.
4. List the merits and demerits of a synchronous detector.
5. Is it possible to excite a tuned circuit with the aid of a parametric capacitor whose capacitance varies in time at a frequency equal to the resonant frequency of the tuned circuit?
6. Describe the physical principle underlying the operation of the varactor.
7. Name the phenomena accompanying the operation of a single-stage parametric amplifier in the asynchronous mode.
8. What is the advantage of double-stage parametric amplifiers?
9. Why is it that parametric amplifiers have a low level of internal noise?
10. Formulate the condition for the "slow" passage of a parametric tuned circuit through resonance.
11. Give an example of a physical system in which the input signal is subjected to random amplitude modulation.
12. List the salient features of the signal spectrum produced by random angle modulation of a harmonic carrier.

Problems

1. The parametric conductance is varying in time as

$$G(t) = 10^{-3} + 5 \times 10^{-4} \cos 10^5 t + 3 \times \times 10^{-4} \times \cos 2 \times 10^5 t$$

The excitation is a voltage

$$v(t) = 5 \cos 10^6 t$$

Find the amplitudes and frequencies of all the current components. Plot the spectral diagram.

2. A lag-free nonlinear resistor has the current-voltage characteristic (mA)

$$i(v) = 5 + 2.5v + 1.5v^2$$

The voltage (V) applied to the resistor is

$$v(t) = 3 + 0.5 \cos \Omega t$$

Derive the equation defining the time dependence of the dynamic transconductance.

3. There is a frequency converter built around a transistor whose characteristic is given by

$$i_C = \begin{cases} 20(v_{BE} - 0.5) & \text{for } v_{BE} > 0.5\ \text{V} \\ 0 & \text{for } v_{BE} < 0.5\text{V} \end{cases}$$

In the "no-signal" condition, the base is fed with the sum of the bias voltage and the local-oscillator voltage (V):

$$v_{BE} = 0.2 + 0.7 \cos \omega_{lo} t$$

Find the conversion transconductance.

4. The capacitance of a parametric capacitor (pF) varies in time as

$$C(t) = 200 + 80 \cos (10^5 t + \pi/4) + 40 \cos 5 \times 10^5 t$$

The voltage applied to the capacitor (V) is

$$v = 30 \cos 5 \times 10^6 t$$

Find an analytic expression for the current in the capacitor.

5. The inductance of a tuned circuit is 0.5 mH, the average capacitance is $C_0 = 750$ pF, the tuned-circuit loss resistance is 12 Ω. The tuned-circuit capacitance varies stepwise by the same amount in either direction from the mean value. Find the frequency at which the tuned-circuit capacitance should be varied and between what limits for the resultant Q of the tuned circuit to be $Q = 300$.

6. Find the Q-factor of a tuned circuit whose inductance is 100 μH and whose loss resistance is 15 Ω. The tuned-circuit capacitance (pF) is varied in time as

$$C(t) = 150 + 5 \cos 1.63 \times 10^7 t$$

7. The capacitance (pF) of a parametric capacitor connected in a tuned circuit is varied in time as

$$C(t) = 300 + 20 \cos 5 \times 10^6 t$$

Find the inductance and the Q-factor at which the system will be excited parametrically and the amplitude of oscillation will build up without bound. Is the solution thus found the only one?

8. A single-stage parametric amplifier has been built to amplify at a frequency of 120 MHz. The amplifier contains an inductor of 0.6 μH; the tuned-circuit Q-factor is 35. Find the frequency at which the tuned-circuit capacitor should be varied and between what limits for the gain of the system to be 15 dB.

Advanced Problems

9. Derive the expression defining time variations in the free response of a parametric RC-network whose capacitance varies in time as

$$C(t) = C_0 + C_m \cos \omega_c t$$

where ω_c is the frequency of the control signal.

10. Find the impulse response of a parametric RC-network whose capacitance varies in time as

$$C(t) = C_0 + C_m \exp(-t/\tau)\,\sigma(t)$$

where C_0, C_m, and τ are constants.

11. A source of d.c. voltage V_0 is connected at time $t = 0$ to a series RC-network for which $R(t) = R_0 \exp(\alpha t)$ and $C(t) = C_0 \exp(-\alpha t)$. Derive equations defining variations in v_R and v_C. Analyse the cases for $\alpha \gg 1/R_0C_0$ and $\alpha \approx 1/R_0C_0$.

Chapter 13

A Basic Theory of Linear Circuit Synthesis

Circuit theory may be divided into two broad areas closely related to each other, namely *analysis* and *synthesis*. The objective of analysis is to find the external characteristics of a system whose structure is specified in advance by giving its schematic diagram. The objective of synthesis is diametrically opposite. We are given an external characteristic, say, the frequency response function, and we are to find the circuit structure implementing the given characteristic.

Circuit synthesis yields more than one result

In contrast to analysis, circuit synthesis will usually yield more than one result. Therefore, it is additionally required to find, from among the several structures possessing the same specified properties, that which is optimal in a particular sense. The criteria of optimality may be many and diverse. For one thing, it is always desired that the circuit being synthesized should contain the least possible count of circuit components. In other cases, it may be desired that the circuit be sensitive only slightly to the choice of element values.

13.1 Analytical Properties of the Driving-Point Impedance of a Passive Linear One-Port

Recently, circuit synthesis methods have come to play an especially important role with the advent of computer-aided circuit design systems. A whole range of synthesis procedures, sometimes extremely sophisticated, have been developed, with which the reader may acquaint himself on his own [27, 30]. This Chapter will be concerned with two very simple problems of synthesis, namely the structures of linear one-ports and two-ports made up of elements such as R, L, and C. In all cases, the basic data for the synthesis will be stated in terms of frequency characteristics.

It is to be noted that the synthesis procedures set forth here can be applied not only to electric or electronic circuits, but also to any linear systems which can be modelled as circuits.

For a network synthesis procedure to be informative, we should above all establish the criteria that could be used to predict the realizability of a network. This section will be concerned with the most important properties of the driving-point impedance of passive linear one-ports.

Location of poles and zeros. Let $V(p)$ and $I(p)$ be the Laplace transforms of the voltage across and the current in a one-port.

Their ratio

$$Z(p) = V(p)/I(p)$$

is the driving-point impedance defined over the entire complex-frequency plane. Also, as can be recalled from Chap. 8,

$$Z(p) = \frac{a_m p^m + a_{m-1}p^{m-1} + \ldots + a_1 p + a_0}{b_n p^n + b_{n-1}p^{n-1} + \ldots + b_1 p + b_0} = \frac{Z_0(p-z_1)(p-z_2)\ldots(p-z_m)}{(p-p_1)(p-p_2)\ldots(p-p_n)} \tag{13.1}$$

Here Z_0 is an arbitrary scale factor

The poles $p_1, p_2, \ldots, p_n$ and the zeroes $z_1, z_2, \ldots, z_m$ of the driving-point impedance of a one-port should be such that the system in question which, by the statement of the problem, does not contain any sources continuously feeding energy, will be stable.

Let the terminals of the one-port be open (the open-circuit or no-load condition) and a voltage exist between them. Since no current is flowing through the one-port, the behaviour of the one-port can be described by the *characteristic equation*

The characteristic equation of a one-port

$$I(p)/V(p) = 1/Z(p) = 0 \tag{13.2}$$

whose roots are the poles of the driving-point impedance $Z(p)$. Any voltage between the terminals of a one-port which is in the open-circuit condition and is storing an amount of energy in its reactive elements is described by

$$v_{o.c}(t) = A_1 \exp(p_1 t) + A_2 \exp(p_2 t) + \ldots + A_n \exp(p_n t)$$

where the A_i's are the coefficients found from the initial conditions. The condition for an open-circuited system to be stable is defined by the inequality

$$\mathrm{Re}(p_i) < 0, \quad i = 1, 2, \ldots, n \tag{13.3}$$

Similarly, if we examine the same one-port with its terminals short-circuited (the short-circuit condition), when $V(p) = 0$, but $I(p) \neq 0$, a second characteristic equation can be derived

$$V(p)/I(p) = Z(p) = 0 \tag{13.4}$$

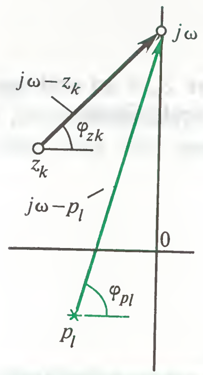
Its roots are the zeros of the driving-point impedance; they should be such that

$$\mathrm{Re}(z_i) < 0, \quad i = 1, 2, \ldots, m \tag{13.5}$$

Location of the poles and zeros of a stable one-port

Thus, both the poles and the zeros of the driving-point impedance of an absolutely stable passive linear one-port *lie solely in the left-hand half of the complex p-plane.* Also, the roots are always either real or form complex-conjugate pairs.

The limiting idealized case is a purely reactive one-port. Since it is free from ohmic losses, its poles and zeros are always located on the imaginary $j\omega$-axis.



The number of poles and zeros. Further information about the behaviour of the function $Z(p)$ is given by the following important theorem: *The number of poles for the driving-point impedance of a passive one-port cannot differ from that of zeros by more than one.*

By way of proof, let us write the driving-point impedance at some physical frequency ω as

$$Z(j\omega) = |Z(j\omega)| \exp(j \arg Z) \tag{13.6}$$

So that Eq. (13.6) can be depicted graphically, we should put $p = j\omega$ in Eq. (13.1) and draw vectors from all poles and zeros so that their tips converge to the selected point on the imaginary axis, representing the current frequency. Then, as can readily be seen, the phase angle of the driving-point impedance is

$$\arg Z = \sum_{k=1}^{m} \varphi_{zk} - \sum_{l=1}^{n} \varphi_{pl}$$

which means that the poles decrease, and the zeros increase the resultant phase.

On the average, a stable passive one-port always draws power from external sources at any frequency. This implies that the real part of the driving-point impedance, $Z(j\omega)$, is positive. Hence,

$$-\pi/2 \leqslant \arg Z \leqslant \pi/2 \tag{13.7}$$

Let the frequency ω tend to infinity. Then both the phases of the poles and the phases of the zeros will tend towards $\pi/2$. Thus,

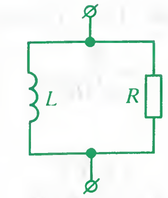
$$\lim_{\omega \to \infty} \arg Z = (m-n)\pi/2 \tag{13.8}$$

Therefore, the numbers m and n may be either the same or differ by unity.

The positive real (p.r.) function

In circuit theory, the function $Z(p)$, analytic in the right-half plane and having a nonnegative real part on the imaginary $j\omega$-axis belongs to a special class of *positive real (p.r.) functions.*

More tangibly the above theorem may be stated like this: As the frequency tends to infinity, any passive network can behave either as a resistor, if the degree of the numerator in Eq. (13.1) is the same as that of the denominator, or as a capacitor, if the degree of the denominator is greater by one than that of the numerator, or, finally, as an inductor if the converse is true.



Relation between the real and imaginary parts of the driving-point impedance. It is an easy matter to verify that at real frequencies ω the driving-point impedance of a one-port made up of a resistor R and an inductor L connected in parallel is defined by

$$Z(j\omega) = R(j\omega) + jX(j\omega) = \frac{\omega^2 R L^2}{R^2 + \omega^2 L^2} + j\frac{\omega L R^2}{R^2 + \omega^2 L^2}$$

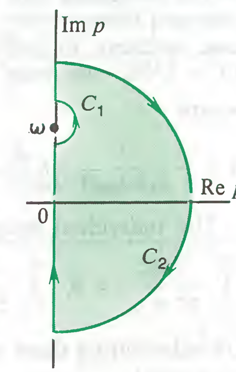
It is to be noted that the values of R and L enter both the real and the imaginary part of the driving-point impedance.

Minimum-impedance networks

Networks composed so that each of their elements affects both the real and the imaginary part of $Z(j\omega)$ are usually called *minimum-impedance networks.* For this class of networks, there is a unique relation between $R(j\omega)$ and $X(j\omega)$ for real values of frequency.

Let us fix the frequency ω and investigate the integral

$$\oint \frac{Z(p)}{p - j\omega}\,dp = 0 \tag{13.9}$$



taken over a closed contour in the right half-plane. The pole arising for $p = j\omega$ is passed around along the semi-circle C_1 of small radius. The integral along the infinitely large arc C_2 is negligible because $Z(p)$ is analytic to the right of the imaginary axis. Therefore,

$$\int_{-\infty}^{\infty} \frac{Z(j\xi)\,d\xi}{\xi - \omega} + \int_{C_1} \frac{Z(p)\,dp}{p - j\omega} = 0 \tag{13.10}$$

The integrand $Z(p)/(p - j\omega)$ in the vicinity of the pole $p = j\omega$ tends to infinity uniformly. Therefore by Cauchy's residue theorem, noting that arc C_1 is a half-circle and the residue at the pole is $Z(j\omega)$, we may re-write Eq. (13.10) as

Only half the residue at a pole contributes to the integral

$$\int_{-\infty}^{\infty} \frac{Z(j\xi)\,d\xi}{\xi - \omega} + j\pi Z(j\omega) = 0 \tag{13.11}$$

Hence, on splitting into the real and imaginary parts, we get

Here the integrals are taken in the sense of the principal value

$$X(j\omega) = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{R(j\xi)\,d\xi}{\xi - \omega}$$
$$R(j\omega) = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{X(j\xi)\,d\xi}{\omega - \xi} \tag{13.12}$$

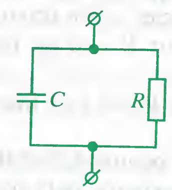
Thus, the real and imaginary parts of the driving-point impedance for the class of one-ports in question are related by a Hilbert pair.

Example 13.1. *There is a parallel RC-network for which*

$$R(j\omega) = \frac{R}{1+\omega^2 (RC)^2} \tag{13.13}$$

and

$$X(j\omega) = \frac{-\omega R^2 C}{1+\omega^2 (RC)^2} \tag{13.14}$$



Verify that, if the real part of the driving-point impedance for this network is known, we can recover the imaginary part by means of Hilbert transformation.

On setting $\alpha = 1/RC$ and inserting (13.13) into (13.12), we get

$$X(j\omega) = \frac{1}{\pi R C^2} \int_{-\infty}^{\infty} \frac{d\xi}{(\alpha^2+\xi^2)(\xi-\omega)} \tag{13.15}$$

The integrand can be expanded into partial fractions:

$$\frac{1}{(\alpha^2+\xi^2)(\xi-\omega)} = \frac{a\xi+b}{\alpha^2+\xi^2} + \frac{c}{\xi-\omega}$$

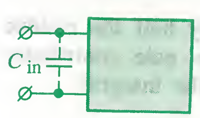
where

$$a = -\frac{1}{\alpha^2+\omega^2}, \quad b = -\frac{\omega}{\alpha^2+\omega^2}, \quad c = -a$$

The individual integrals have the form

$$\int_{-\infty}^{\infty} \frac{\xi\, d\xi}{\alpha^2+\xi^2} = 0, \quad \int_{-\infty}^{\infty} \frac{d\xi}{\xi-\omega} = 0, \quad \int_{-\infty}^{\infty} \frac{d\xi}{\alpha^2+\xi^2} = \pi/\alpha$$

On substituting them in (13.15), we obtain Eq. (13.14), which was to be proved.



The result thus obtained is very important in assessing the behaviour of the driving-point impedance of a one-port along the entire frequency axis. As an example, we may quote a theorem [36] which establishes that the driving-point impedance of any network, such as an amplifier, whose driving-point capacitance is C_{in}, satisfies the inequality

Solve Problem 1

$$\int_{0}^{\infty} R_{in}\, d\omega \leqslant (\pi/2)/C_{in} \tag{13.16}$$

On this ground we may state, for example, that an amplifier with a flat driving-point impedance $R_{in} = 1$ MΩ over the frequency interval 0-1 MHz, with the input capacitance being $C_{in} = 10$ pF, is unrealizable in principle.

The driving-point impedance of reactive one-ports. The frequency properties of the driving-point impedance of purely reactive

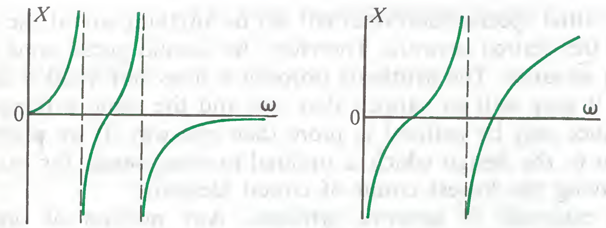


Fig. 13.1 Typical plots of functions $X(j\omega)$ for two purely reactive networks

one-ports are concern of *Foster's theorem* [27], an important proposition in circuit theory. It is stated as follows: If $Z(j\omega) = jX(j\omega)$, then the reactive impedance is a nondecreasing function, that is, $dX/d\omega > 0$. The frequency dependence of the driving-point impedance for some reactive one-ports is plotted in Fig. 13.1.

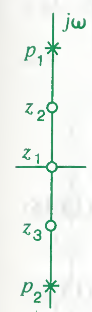
Foster's theorem

As a corollary to the Foster theorem, we may state that (a) points $p = 0$ and $p = \infty$ are the singularities (zeros or poles) of the driving-point impedance; (b) on the $j\omega$-axis, the poles and zeros alternate.

The properties of the driving-point impedances of reactive one-ports

In synthesizing an electric network, it is mandatory first to analyse the function $Z(p)$ for its properties so as to discard the obviously unrealizable characteristics.

Work Problem 2



Example 13.2. *We are given the function*

$$Z(p) = \frac{p(p^2+1)}{p^2+4} \tag{13.17}$$

which has three zeros: $z_1 = 0$, $z_2 = j$ *and* $z_3 = -j$; *and two poles:* $p_1 = 2j$ *and* $p_2 = -2j$.

The singularities are located on the imaginary axis (which is typical of reactive networks). The number of poles is by one smaller than the number of zeros. Nevertheless the network described by the function in (13.17) cannot be realized because it violates the requirement that the singularities should alternate on the $j\omega$-axis.

13.2 Synthesis of Passive One-Ports

The classical problem of one-port synthesis can be stated as follows: We are given the driving-point immitance of the desired network, which may be its driving-point impedance $Z(p)$ or its driving-point admittance $Y(p)$, that meets all the requirements placing it among the functions of physically realizable and stable one-ports. We are to synthesize the network answering the specified driving-point immitance.

The problem of one-port synthesis

The initial specification does not tell us anything about the structure of the desired network. Therefore, we should guess some structure in advance. The synthesis procedure does not yield a unique result: it may well so happen that one and the same driving-point immitance may be realized in more than one way. If so, preference is given to the design which is optimal in some sense, for example, one having the lowest count of circuit elements.

The rationale of network synthesis. Any method of one-port synthesis is based on the fact that the specified driving-point immitance, $Z(p)$ or $Y(p)$, is subjected to a series of consecutive simplifications. At each step, we single out a particular expression which can be uniquely identified with a particular circuit element. The nature of these transformations is fixed in advance by the network structure we have chosen. The simplest structures we are going to consider are shown in Fig. 13.2.

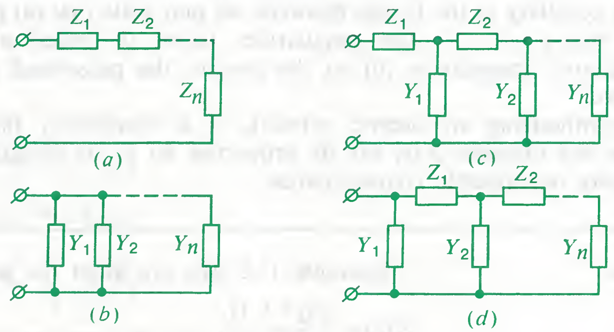


Fig. 13.2 Some of the one-ports being synthesized

It is an easy matter to see that in case (*a*)

$$Z(p) = Z_1 + Z_2 + \ldots + Z_n \tag{13.18}$$

and in case (*b*)

$$Y(p) = Y_1 + Y_2 + \ldots + Y_n \tag{13.19}$$

Ladder networks

In practical synthesis problems, the number of structure elements is always finite

The one-ports shown in Fig. 13.2*c* and *d* are known as *ladder networks*. Here, for case (*c*)

$$Z(p) = Z_1 + \cfrac{1}{Y_1 + \cfrac{1}{Z_2 + \cdots \cfrac{1}{Z_n + \cfrac{1}{Y_n}}}} \tag{13.20}$$

and for case (*d*)

$$Y(p) = Y_1 + \cfrac{1}{Z_1 + \cfrac{1}{Y_2 + \cdots \cfrac{1}{Z_n + \cfrac{1}{Y_n}}}} \tag{13.21}$$

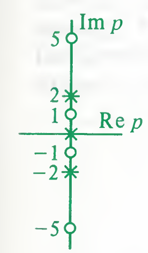
The expressions in (13.20) and (13.21) are called *continued fraction expansions*.

Continued fraction expansion

By identifying the components $Z_1, Z_2, \ldots, Z_n$ and $Y_1, Y_2, \ldots, Y_n$ with specific physical elements we solve the task of one-port synthesis.

Foster and Cauer networks

Synthesis of reactive one-ports. We will dwell upon the basic techniques used in the synthesis of linear passive one-ports, taking as an example purely reactive networks made up of L and C. For this case, it is rigorously shown [27] that any realizable driving-point immitance, $Z(p)$ or $Y(p)$, may be represented as the driving-point immitance of the networks shown in Fig. 13.2. The reactive one-ports shown in (*a*) and (*b*) are known as *Foster type networks*, and those in (*c*) and (*d*), are called *Cauer ladder networks*. Let us turn to a concrete example.



Example 13.3. *The driving-point immitance is given as*

$$Z(p) = \frac{(p^2+1)(p^2+25)}{p(p^2+4)} = \frac{p^4+26p^2+25}{p^3+4p} \tag{13.22}$$

Synthesize a network having the above driving-point impedance by the Foster (or partial fraction expansion) method.

The direct division of the numerator into the denominator in (13.22) identifies the first network element:

$$Z(p) = p + \frac{22p^2+25}{p^3+4p} = p + Z'(p)$$

which is an inductance of 1 H. (Here and elsewhere the network element values are chosen to make the calculations easy to follow.)

The function $Z'(p)$ can be simplified by expanding it into partial fractions:

$$Z'(p) = \frac{25}{4p} + \frac{63p/4}{p^2+4} = \frac{25}{4p} + Z''(p)$$

The first term on the right-hand side represents a capacitor of $4/25 = 0.16$ F capacitance.

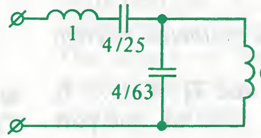
In order to identify the last network element associated with

$Z''(p)$, it is advantageous to change to the driving-point admittance

$$Y''(p) = 1/Z''(p) = \frac{p^2+4}{63p/4} = \frac{4}{63}p + \frac{16}{63p}$$

which represents a parallel combination of a capacitor of 4/63 F capacitance and an inductor of 63/16 H inductance.

The final step is to draw up a schematic diagram which implements the specified driving-point impedance:



▲ **Solve Problem 3**

In order to synthesize a reactive one-port by the Cauer (or continued-fraction expansion) method, the specified driving-point immitance should be expanded into a continued fraction series. The procedure leading to this form stems from the very structure of the fraction. It is called the "*divide and invert the remainder*" process.

Example 13.4. *Synthesize a network having the driving-point impedance defined in* (13.22) *by the Cauer (continued-fraction expansion) method.*

We divide the numerator into the denominator, beginning with the highest degree:

$$\begin{array}{r|l} p^4+26p^2+25 & p^3+4p \\ \cline{2-2} -\ p^4+4p^2 \quad & p \\ \hline 22p^2+25 & \end{array} \tag{13.23}$$

The quotient is the first network element which, as in the Foster type network, is a series inductance of 1 H.

On writing

$$Z(p) = p + \cfrac{1}{\cfrac{p^3+4p}{22p^2+25}} \tag{13.24}$$

we see that in (13.23) the denominator must be divided into the remainder—this is the inversion of the remainder. From this step on, division and inversion alternate:

▲ **Work Problem 4**

Divide:

$$\begin{array}{r|l} p^3+4p & 22p^2+25 \\ \cline{2-2} -\ p^3+\frac{25p}{22} & \frac{p}{22} \\ \hline \frac{63p}{22} & \end{array}$$

Invert the remainder

$$\begin{array}{r|l} 22p^2+25 & \frac{63p}{22} \\ \cline{2-2} -\ 22p^2 \quad & \frac{484p}{63} \\ \hline 25 & \end{array}$$

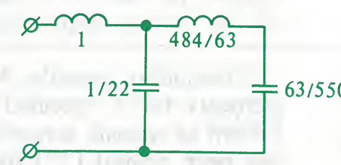
Divide

$$\begin{array}{r|l} \frac{63p}{22} & 25 \\ \cline{2-2} & \frac{63p}{550} \end{array}$$

Combining the results, we can write the continued fraction expansion for $Z(p)$ as

$$Z(p) = p + \cfrac{1}{\cfrac{p}{22} + \cfrac{1}{\cfrac{484}{63}p + \cfrac{1}{\cfrac{63p}{550}}}} \tag{13.25}$$

which can be identified with the following Cauer ladder network:



The two previous examples demonstrate that one-port synthesis does not yield a unique result. In fact, we can synthesize one more network having the same driving-point impedance by the Cauer (continued-fraction expansion) method, if we begin division with the lowest, rather than the highest degrees of the polynomials.

Example 13.5. *Synthesize a network whose driving-point impedance is defined by* (13.22), *using the Cauer (continued-fraction expansion) method and beginning division with the lowest degrees of the polynomials.*

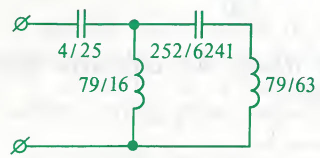
On writing

$$Z(p) = \frac{25+26p^2+p^4}{4p+p^3}$$

dividing and inverting the remainder, we get

$$Z(p) = \frac{25}{4p} + \cfrac{1}{\cfrac{16}{79p} + \cfrac{1}{\cfrac{6241}{252p} + \cfrac{79p}{63}}} \tag{13.26}$$

The above continued-fraction expansion gives us a reactive one-port arranged as shown below:



The Cauer method involving division and inversion of the remainder is convenient for computer-aided network synthesis. It permits a large number of variant techniques. For example, having begun the division with the highest degrees, we may reverse the order of powers during the subsequent steps.

Canonic networks

Concluding remarks. Networks giving the minimum number of elements for a specified driving-point immitance are sometimes known as *canonic networks*. With reference to reactive one-ports it has been proved [27] that both the Foster and the Cauer forms possess this property. If N is the number of pairs of singularities (poles and zeros) of the driving-point immitance, other than zero or infinity, a canonic network contains $N+1$ reactive elements.

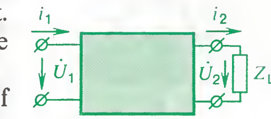
In practical synthesis it should be remembered that the expansion of $Z(p)$ or $Y(p)$ into a sum of partial fractions or continued fractions may produce negative coefficients. This implies that the adopted procedure leads to a nonrealizable network and that a different technique should be used. In the Cauer method, for example, it may prove advisable to begin division with the lowest rather than the highest degrees. It may so happen that in synthesizing an *RLC*-network of the general form none of the methods presented here can yield an acceptable result. This implies that the driving-point immitance involved cannot be realized in any of the proposed simple structures and recourse must be had to more complex forms such as described in [27]. What we may only say is that the desired network does exist, because, as follows from *Darlington's theorem* (of fundamental significance to circuit theory) [28], any physically allowable driving-point function can be realized as the driving-point impedance or admittance of some purely reactive passive four-terminal (two-port) network loaded into a single resistor.

Darlington's theorem

The search for suitable structures is the most crucial step in network synthesis. It commands special attention when developing an algorithm for computer-aided network synthesis.

13.3 Frequency Characteristics of Two-Ports

A four-terminal (or two-port) network may be visualized as a black box with two pairs of accessible terminals (two ports), one pair (or port) being its input, and the other pair (or port) its output. In operation, the input port is connected to a signal source, and the output port is connected to a load impedance Z_L.



It is presumed that the reader has a previous knowledge of two-port analysis methods as they are explained in a course on circuit theory. This section will be concerned only with some points essential to two-port synthesis.

Matrix notation. A very important property of a linear stationary two-port is that the four complex amplitudes $\dot{V}_1$, $\dot{I}_1$, $\dot{V}_2$ and $\dot{I}_2$ are connected by two linear algebraic equations at any frequency of the excitation. If we choose any two of them as independent variables, the remaining two will be defined in their terms. This provides a basis for the matrix description of linear two-ports [25]. For example, one frequently uses the transmission (transfer or *ABCD*) matrix, taking the output voltage and the output current as the independent variables. Then,

$$\dot{V}_1 = A\dot{V}_2 + B\dot{I}_2$$
$$\dot{I}_1 = C\dot{V}_2 + D\dot{I}_2 \tag{13.27}$$

The coefficients A, B, C, and D have different physical dimensions and can be determined from open-circuit and short-circuit tests. Transmission matrices are especially convenient in describing a cascade connection of two-ports, because the resultant matrix is the product of the individual matrices.

Network functions

If the matrix of a two-port and its load impedance are known in advance, we can find the so-called *network functions* such as

(a) the driving-point impedance $Z_{in} = \dot{V}_1/\dot{I}_1$;

(b) the transfer impedance $Z_t = \dot{V}_2/\dot{I}_1$;

(c) the voltage-ratio transfer, or frequency response, function $K(j\omega) = \dot{V}_2/\dot{V}_1$.

In the general case, the network functions are frequency-dependent and characterize the frequency properties of two-ports in a variety of problems. Any network function can be expressed in terms of the members of the two-port's transmission matrix and in terms of the load impedance. Thus, by dividing the respective sides of Eqs. (13.27), we find that

$$Z_{in} = (AZ_L + B)/(CZ_L + D) \tag{13.28}$$

Similarly, the voltage-ratio transfer function is found to be

$$K(j\omega) = \dot{V}_2/\dot{V}_1 = Z_L/(AZ_L + B) \tag{13.29}$$

It is to be noted that $K(j\omega)$ depends on the direction in which power is transferred in the network. If we interchange the source and the load, we should write the voltage-ratio transfer function for the reverse direction as

$$K_{\mathrm{rev}}(j\omega) = \dot{V}_1/\dot{V}_2 \,|\, Z_{L\,\text{on the left}} \tag{13.30}$$

In the general case, the forward and reverse voltage-ratio transfer functions are not the same.

The transfer function of a two-port. In our further discussion, the argument of the transfer function will be not only the variable $j\omega$, but also the complex frequency p. In other words, we shall use both $K(j\omega)$ and $K(p)$, the latter being the more general form of the transfer function. The transfer function of a two-port has all the properties of the same functions of linear stationary systems examined in Chap. 8. For example, the transfer function of a linear constant-parameter two-port has the form

$$K(p) = K_0 \frac{(p-z_1)(p-z_2)\ldots(p-z_m)}{(p-p_1)(p-p_2)\ldots(p-p_n)} \tag{13.31}$$

Location of the poles of the transfer function of a two-port

where K_0 is a constant. If the network is stable, then the poles p_1, p_2, ..., p_n must be located in the left half-plane, occurring as complex-conjugate pairs.

When the transfer function has a zero at infinity, the amplitude response shows a high rate of roll-off at high frequencies

Usually, an additional condition is introduced: *The number of poles of $K(p)$ must be greater than that of zeros*, which means that at infinity the transfer function must have a zero rather than a pole. In consequence, the impulse response of the network

$$h(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} K(j\omega)\exp(j\omega t)\,d\omega = \frac{1}{2\pi j} \int_C K(p)\exp(pt)\,dp$$

must be bounded, because with an infinitely large radius of the integration contour C the exponential term of the integrand would "cancel" the integral around the arc.

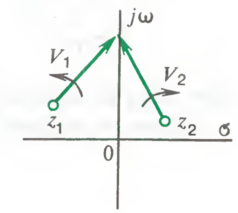
In contrast to what we have for the driving-point impedance of a one-port, the difference in number between the poles and zeros of the transfer function may be any. This is because no energy constraint may be imposed on the phase response.

Location of the zeros of the transfer function. In contrast to the poles, the zeros of $K(p)$ for a stable linear two-port may be located in both the left and right half of the complex p-plane. This is because the characteristic equation $K(p) = 0$ signifies that, given some $V_1(p) \neq 0$, the Laplace transform of the output voltage, $V_2(p)$, vanishes. This does not run counter to the assumption that the system is stable.

Minimum-phase networks

Two-port networks which have no zeros in the right half-plane are called *minimum-phase networks*. If they do have zeros in the right half-plane, they are called *nonminimum-phase networks*.

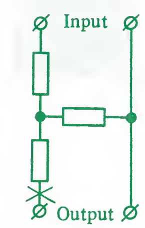
The above names owe their origin to the following. Let there be a complex plane with points z_1 and z_2 in the left and the right half-plane. We also let these points be the zeros of the transfer function of a two-port. If the two-port is driven by a harmonic excitation such that $p = j\omega$, then the two points correspond to two phasors on the complex plane, namely: $V_1 = j\omega - z_1$ and $V_2 = j\omega - z_2$, corresponding to the respective factors in the numerator of Eq. (13.31). The two phasors rotate and vary in length with changes in the frequency ω, but they do so differently. The difference is that, as the frequency varies from $-\infty$ to $+\infty$, the phasor V_1 increases the phase angle of the voltage-ratio transfer function by π radians, whereas the phasor V_2 decreases the phase by the same amount. The voltage-ratio transfer function of a two-port is a rational function whose argument varies as



$$\Delta \arg K(j\omega)\Big|_{\omega=-\infty}^{\omega=+\infty} = \Delta\arg(\text{numer.}) - \Delta\arg(\text{denomin.})$$

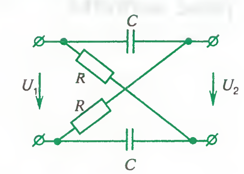
Therefore, given the same number of zeros and poles, a nonminimum-phase network will show a larger change in the phase of the transfer function than a minimum-phase network.

The location of zeros for $K(p)$ is related to the topological structure of the network. It is shown in circuit theory that any two-port will be a minimum-phase network if it has the property such that the transmission of the signal from input to output can be completely discontinued by breaking only one branch. Notably, any ladder-type two-ports are minimum-phase networks.



As a rule, nonminimum-phase networks are *bridge* and *lattice networks* in which the signal can reach the output over two and more paths. The simplest example of a nonminimum-phase network is a balanced lattice network formed by R and C. Here, as can be seen,

$$K(p) = (pRC-1)/(pRC+1) \tag{13.32}$$



This transfer function has a zero at point $p = 1/RC$, that is, in the right half-plane.

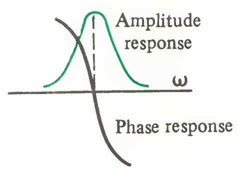
However, the bridge or lattice structure does not automatically guarantee that the network is a minimum-phase one, so it is essential to verify the presence or otherwise of zeros in the right half-plane in each particular case.

Relation between the magnitude and phase responses of a two-port. It is proved [1] that minimum-phase two-ports have a remarkable property: Their magnitude and phase responses are uniquely related to each other. The real and imaginary parts of the logarithm of the frequency response function

$$\ln\{|K(j\omega)|\exp[j\varphi_K(\omega)]\} = \ln|K(j\omega)| + j\varphi_K(\omega) = \psi_K(\omega) + j\varphi_K(\omega)$$

form a Hilbert transform pair

$$\varphi_K(\omega) = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{\psi_K(\xi)\,d\xi}{\xi - \omega}$$
$$\psi_K(\omega) = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{\varphi_K(\xi)\,d\xi}{\omega - \xi} \tag{13.33}$$



A minimum-phase two-port realizing a specified magnitude response cannot have just any phase response. Basing ourselves on the properties of the Hilbert transforms (see Chap. 5), we may argue, for example, that if the magnitude response of a minimum-phase two-port is a maximum at some particular frequency, its phase response will cross zero in the vicinity of that frequency.

Uses of nonminimum-phase networks

If, on the other hand, a two-port belongs to the nonminimum-phase class, its magnitude and phase responses are independent of each other. Among the nonminimum-phase networks, the most important role is played by all-pass two-port networks whose magnitude response is constant and independent of frequency. An example is a balanced lattice RC two-port for which (see (13.32))

$$|K(j\omega)| = 1, \quad \varphi_K = -2\arctan\omega RC \tag{13.34}$$

Such two-ports are used for the phase compensation of signals. They are able to make up in part for the distortion of signals passing through a circuit.

The power transfer function. As will be recalled (see Chap. 8), this term refers to the square of the magnitude response of a two-port:

$$K_P(\omega) = K(j\omega)K^*(j\omega) = K(j\omega)K(-j\omega) \tag{13.35}$$

In contrast to $K(j\omega)$, the power transfer function $K_P(\omega)$ is real. Because of this it is especially convenient to state the initial specifications needed to synthesize a two-port in its terms. Unfortunately, it does not tell us anything about the phase response of the system.

The frequency dependence of the power transfer function

As is seen from (13.35), the power transfer function is an even function of frequency, so it can always be represented as the ratio of two polynomials in powers of ω^2:

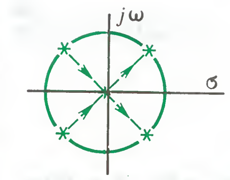
$$K_P(\omega) = M(\omega^2)/N(\omega^2) \tag{13.36}$$

By a change of variable $p = j\omega$, the power transfer function $K_P(\omega)$ is continued analytically from the imaginary $j\omega$-axis over the entire complex frequency plane:

$$K_P(p) = K(p)K(-p) \tag{13.37}$$

Equation (13.37) establishes a very important fact: If $a + jb$ is

a singularity (a zero or a pole) of the function $K(p)$, then $K_P(p)$ will likewise have a singularity for both $p = a + jb$ and for $p = -a - jb$. It is customary to say that the singularities of the power transfer function show *quadrant symmetry*, that is, they are located in the complex plane so that their centre of symmetry is at the origin. This property is of special value for two-port synthesis, as it permits one to recover the frequency response function from a known $K_P(p)$.



Location of poles with quadrant symmetry

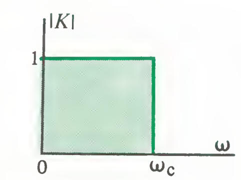
13.4 Low-Pass Filters

This section will be concerned with the frequency behaviour of low-pass filters whose function is to transmit with a minimal attenuation all frequencies from zero up to a desired *cut-off frequency*, ω_c, and to attenuate all higher frequencies (from ω_c to infinity).

The cut-off frequency of a filter

Steps in filter synthesis. As a rule, the first step in the synthesis of frequency-selective networks is to formulate the requirements for their frequency response. For example in the case of a low-pass filter with a cut-off frequency ω_c the ideal magnitude response has the form

$$|K(j\omega)| = \begin{cases} 1, & 0 \leqslant \omega \leqslant \omega_c \\ 0, & \omega > \omega_c \end{cases} \tag{13.38}$$



(physical frequencies, $\omega > 0$, are meant). No constraints are imposed on the phase response. This approach is known as the synthesis of a filter from the specified magnitude response.

It is known in advance that the ideal magnitude response defined in (13.38) is not physically realizable (see Chap. 8). Therefore, the second step of filter synthesis is to approximate the ideal characteristic with a function which may belong to a physically realizable network.

The final step is to realize the selected characteristic and to develop a circuit diagram complete with all the necessary element values.

The two forms of approximation most frequently used in practical filter design are the *maximally flat* (or *Butterworth*) *approximation* and the *equal-ripple* (or *Chebyshev*) *approximation*.

The maximally flat approximation. This form of approximation is based on the use of the power transfer function

In the general case, the power transfer function may contain an arbitrary scale factor

$$K_P(\omega_N) = 1/(1 + \omega_N^{2n}) \tag{13.39}$$

where $\omega_N = \omega/\omega_c$ is a dimensionless normalized frequency. Low-pass filters having such a characteristic are said to have

a *maximally flat* or *Butterworth response*, and the filters themselves are called *maximally flat* or *Butterworth filters*. The integer $n = 1, 2, 3, \ldots$ controls the closeness of approximation and is called the *order of a filter*. From comparison of (13.36) and (13.39) it can be seen that a Butterworth filter is realizable for any value of n. Within the pass band, that is, for $0 \leqslant \omega_N \leqslant 1$, the square of the magnitude response asymptotically decreases with increasing frequency. At the cut-off frequency (for $\omega_N = 1$) the asymptotic slope produced by the filter is $10 \log_{10} 0.5 \approx -3$ dB, irrespective of the filter's order. The greater the value of n, the closer the approximation. The maximally flat response curves for two values of n, constructed on the basis of Eq. (13.39), appear in Fig. 13.3.

● **The order of a filter**

▲ **Solve Problem 5**

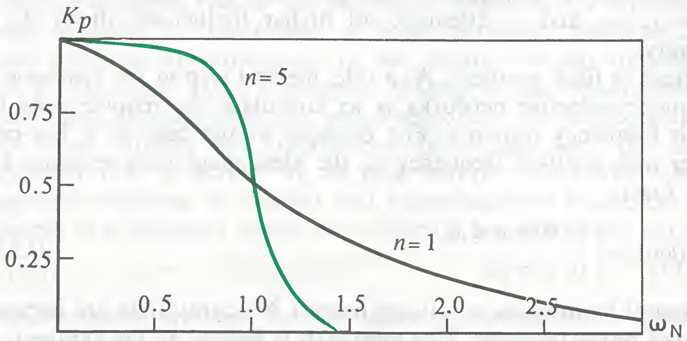


Fig. 13.3 Frequency dependence of the power transfer function of $n = 1$ and $n = 5$ Butterworth filters

The parameter n is chosen according to the degree of attenuation that should be given to signals at frequencies $\omega > \omega_c$.

Example 13.6. *Find the parameter n for a Butterworth filter with a cut-off frequency of 10^5 s^{-1}, whose attenuation at $\omega = 3 \times 10^5$ s^{-1} would be at least -26 dB relative to the level at $\omega = 0$.*

From the statement of the problem, we conclude that the parameter n must be the nearest integer (with an excess) to the solution of the equation

$$10 \log_{10} \frac{1}{1 + 3^{2n}} = -26$$

or

$$1 + 3^{2n} = 10^{2.6} = 398$$

On solving it, we get

$$2n = \log_{10} 397 / \log_{10} 3 = 5.45$$

Hence, $n = 3$.

When the signal is appreciably distant from the pass band, then $\omega_N \gg 1$, and from Eq. (13.39) we have

▲ **Work Problem 6**

$$K_P(\omega_N) \approx \omega_N^{-2n}$$

Hence, the asymptotic slope (attenuation) on the decibel scale is

$$\Delta = 10 \log_{10} K_P \approx -20n \log_{10} \omega_N$$

An octave is the interval between two points having a basic frequency ratio of two

This implies that each time the frequency is doubled, the attenuation produced by a Butterworth filter is increased by $-20n \times 0.301 \approx -6n$ dB. That is, the magnitude response falls asymptotically at a rate of $6n$ dB/octave outside the pass band.

The transfer function of a maximally flat (Butterworth) filter. In order to be able to synthesize the filter, we should change from the power transfer function as defined in (13.39) to the transfer function $K(p)$. To this end, we introduce the normalized complex frequency $p_N = \sigma_N + j\omega_N$ and re-write (13.39) as

$$K_P(p_N) = K(p_N) K(-p_N) = \frac{1}{1 + (-1)^n p_N^{2n}} \qquad (13.40)$$

Hence it is seen that in the p_N-plane the function $K_P(p_N)$ answering a Butterworth low-pass filter of the nth order has $2n$ poles which are the roots of the equation

$$1 + (-1)^n p_N^{2n} = 0 \qquad (13.41)$$

All of these roots are located on the unit circle with centre at the origin. For $n = 1$, the coefficients of the power transfer function are found from the equation

$$p_N^2 = 1$$

that is

$$p_{N1} = 1, \quad p_{N2} = -1 \qquad (13.42)$$

For $n = 2$, the equation

$$p_N^4 = -1$$

has four roots:

▲ **Solve Problem 7**

$$p_{N_1} = e^{j\pi/4}, \; p_{N_2} = e^{j3\pi/4}, \; p_{N_3} = e^{j5\pi/4}, \; p_{N_4} = e^{j7\pi/4} \qquad (13.43)$$

Finally, for a third-order ($n = 3$) Butterworth filter, the equation is

$$p_N^6 = 1$$

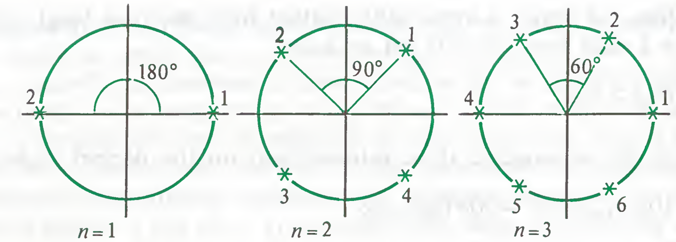
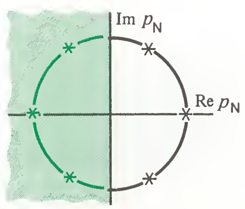


Fig. 13.4 The poles of the power transfer function for a Butterworth low-pass filter with $n = 1, 2, 3$

which has six roots:

$$p_{N_1} = 1,\ p_{N_2} = e^{j\pi/3},\ p_{N_3} = e^{j2\pi/3}$$
$$p_{N_4} = -1,\ p_{N_5} = e^{j4\pi/3},\ p_{N_6} = e^{j5\pi/3} \qquad (13.44)$$



The selection of poles of the transfer function

The location of the roots in the complex plane for the cases listed above is shown in Fig. 13.4.

The general trend for any value of n is this: All poles are spaced apart the same angular distance equal to π/n; if n is an odd number, the first root is $p_{N_1} = 1$; if n is even, then $p_{N_1} = \exp(j\pi/n)$.

Now we will take advantage of the fact that the poles of the power transfer function show quadrant symmetry, that is, their number and location in the two half-planes are the same. Hence, we may rightfully conclude that only the poles located in the left half-plane answer the filter being synthesized. Their "mirror images" in the right half-plane are associated with the function $K(-p)$ and are ignored.

The principle described above is the pivotal one in filter synthesis, because the subsequent realization of the network is based on it.

Example 13.7. *Find the transfer function of an $n = 2$ Butterworth low-pass filter.*

The transfer function has two poles in the left half-plane (see Eq. (13.43)):

$$p_{N_2} = (-1 + j)/\sqrt{2},\ p_{N_3} = (-1 - j)/\sqrt{2}$$

Hence

$$K(p_N) = \frac{1}{(p_N - p_{N_2})(p_N - p_{N_3})} = \frac{1}{p_N^2 + \sqrt{2}\,p_N + 1} \qquad (13.45)$$

Thus, an $n = 2$ Butterworth filter can be implemented with a second-order dynamic system (an oscillatory element).

Chebyshev approximation. Also known as the equal-ripple approximation, this technique is likewise widely used. For the resultant form of filter, the power transfer function is defined as

$$K_P(\omega_N) = \frac{1}{1 + \varepsilon^2 C_n^2(\omega_N)} \qquad (13.46)$$

The ripple factor

where $\varepsilon < 1$ is a constant number called the *ripple factor in the pass band*, and $C_n(\omega_N)$ is the nth-order Chebyshev polynomial* defined by

$$C_n(x) = \cos(n \operatorname{arc\,cos} x) \qquad (13.47)$$

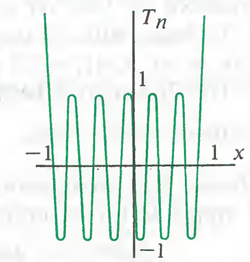
Higher-order Chebyshev polynomials are obtained through the recursive (or recursion) formula

$$C_n(x) = 2xC_{n-1}(x) - C_{n-2}(x) \qquad (13.48)$$

such that $C_0(x) = 1$ and $C_1(x) = x$.

Chebyshev polynomials are frequently used in all kinds of approximation problems owing to the following property: Among any polynomials of degree n with the same coefficients of the highest power of the argument, these polynomials deviate least of all from zero in the interval $-1 < x < 1$. On the other hand, Chebyshev polynomials sharply increase in magnitude at $|x| \gg 1$. Asymptotically,

$$C_n(x) \approx 2^{n-1}x^n \text{ for } |x| \gg 1 \qquad (13.49)$$



A typical plot of a Chebyshev polynomial

The property of Chebyshev polynomials

Such functions are convenient to use in the ideal low-pass filter approximation. As is seen from (13.46), within the pass band K_P oscillates about unity such that the maximum value is 1 and the minimum is $1/(1 + \varepsilon^2)$. Outside the pass band, that is, at $\omega_N \gg 1$, the filter attenuates the signal appreciably.

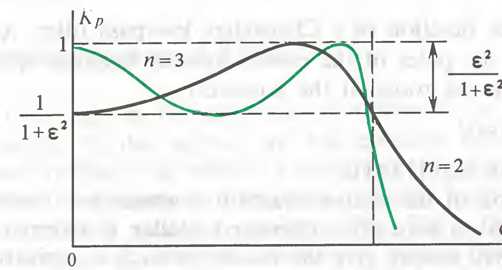


Fig. 13.5 Frequency dependence of the power transfer function of a Chebyshev low-pass filter

* An alternative symbol for Chebyshev polynomials is $T_n(\omega)$ from the older transliteration "Tschebishev".– *Translator's note.*

Typical plots of the power transfer functions for $n = 2$ and $n = 3$ Chebyshev filters appear in Fig. 13.5. As is seen, within the pass band the frequency characteristics of the filters are nonmonotonic. The ripple in the attenuation increases with the value of ε. As follows from (13.46), an increase in ε leads to a greater attenuation of the signal outside the pass band. By matching two variables, n and ε, it is possible to satisfy the initial specifications for the filter being synthesized.

▲ **Solve Problem 8**

Example 13.8. *There is an $n = 3$ Chebyshev filter which, at the cutoff frequency ($\omega_N = 1$), attenuates the signal in power by half, that is, by the same amount as a Butterworth filter. Find the attenuation produced by this filter at a frequency three times the cutoff frequency.*

To begin with, let us find ε. As follows from (13.47), $C_n(1) = 1$ for any n, so $K_P(1) = 1/2$ if $\varepsilon = 1$.

The 3rd-order Chebyshev polynomial is

$$C_3(\omega_N) = 4\omega_N^3 - 3\omega_N$$

Hence, the attenuation produced by this Chebyshev filter with a ripple factor of unity at frequency $\omega = 3\omega_c$ will be

$$\Delta_{Cheb} = 10\log_{10}\frac{1}{1 + C_3^2(3)} = -39.91\ \text{dB}$$

It is interesting to note that under similar conditions an $n = 3$ Butterworth filter will produce an attenuation equal to

$$\Delta_{But} = 10\log_{10}\frac{1}{1 + 3^6} = -28.36\ \text{dB}$$

Thus, a Chebyshev filter appreciably improves the performance of a frequency-selective system.

The transfer function of a Chebyshev low-pass filter. As is seen from (13.46), the poles of the power transfer function of a Chebyshev filter are the roots of the equation

$$1 + \varepsilon^2 C_n^2(p_N) = 0 \tag{13.50}$$

(compare with Eq. (13.41)).

The solution of the above equation is somewhat involved, and will not be given here. The interested reader is referred to [28]. Instead, we will simply give the results of such a derivation. First we introduce a design parameter

$$a = \frac{1}{n}\sinh^{-1}\frac{1}{\varepsilon} = \frac{1}{n}\ln\left(\frac{1}{\varepsilon} + \sqrt{\frac{1}{\varepsilon^2} + 1}\right) \tag{13.51}$$

The next step is to find the poles of a Butterworth filter with the

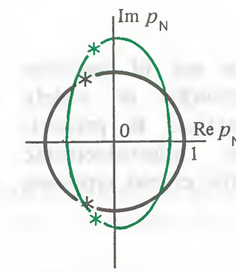
same value of n. Then we can change to the poles of the Chebyshev filter by multiplying the abscissa of each pole of the Butterworth filter by $\sinh a$, and the ordinate by $\cosh a$.

Whereas the poles of a Butterworth filter are located on the unit circle, those of a Chebyshev filter are located on an ellipse which, on the $p_N = \sigma_N + j\omega_N$ plane, is specified by

$$(\sigma_N/\sinh a)^2 + (\omega_N/\cosh a)^2 = 1$$

Once the coordinates of the poles are obtained, we can write the transfer function of the Chebyshev filter as

$$K(p_N) = \frac{1}{(p_N - p_{N1})(p_N - p_{N2})\dots(p_N - p_{Nn})}$$



The poles of the Butterworth and the Chebyshev filters

Example 13.9. *Find the transfer function of an $n = 2$ Chebyshev filter with $\varepsilon = 1$.*

Here

$$a = \frac{1}{2}\ln(1 + \sqrt{2}) = 0.4407$$

▲ **Work Problem 9**

The corresponding Butterworth filter has two poles:

$$p_{N1} = 0.707(-1 + j) \text{ and } p_{N2} = 0.707(-1 - j)$$

Hence, the abscissae of the Chebyshev filter are $-0.707\sinh a = -0.322$, and the ordinates are $\pm 0.707\cosh a = \pm 0.777$.

It is seen from the above example that in going from a maximally flat to a Chebyshev response, the poles move towards the imaginary axis, whereas their vertical displacement is insignificant. Physically this signifies that the oscillatory system making up the Chebyshev filter will produce a smaller attenuation.

13.5 Implementation of Filters

The final step in the synthesis of a filter is to find its circuit arrangement. In this section we will consider what is known as *structural synthesis* in which a circuit is formed by a cascade connection of a number of sections or stages separated from one another by ideal *isolation networks* (Fig. 13.6). The voltage-ratio transfer (frequency response) function of such a device is given by

● **Structural synthesis**

$$K(j\omega) = K_1(j\omega)K_2(j\omega)\dots K_N(j\omega)$$

The individual functions K_1, K_2, ..., K_N must be such that they realize the poles of the transfer function $K(p)$, defined earlier during the approximation step.

The use of isolation networks is widely practised in present-day microelectronic active circuit synthesis

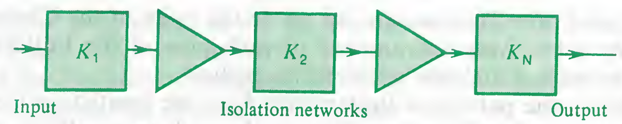
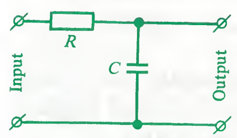


Fig. 13.6 Filter formed by a cascaded connection of elements (the isolation networks are usually emitter followers)

It takes two kinds of networks to build a low-pass filter, namely a 1st-order network which has only one real pole, and a second-order network which has a pair of complex conjugate poles.

The 1st-order network. The simplest example of this kind is an L-section RC two-port for which

$$K(p) = 1/(1 + pRC) \tag{13.52}$$



the coordinate of the pole being $p_1 = -1/RC$.

It is to be noted that by specifying p_1, we only obtain the product RC. One of these two elements, R or C, may be chosen at will. For example, it may be desired that the capacitance C be substantially greater than the input capacitance of the succeeding stage. This will make the filter less sensitive towards inaccuracies in the selection of the element values.

The 2nd-order network. Two complex conjugate poles of the frequency response function can be realized with the aid of the L-section two-port shown schematically in Fig. 13.7.

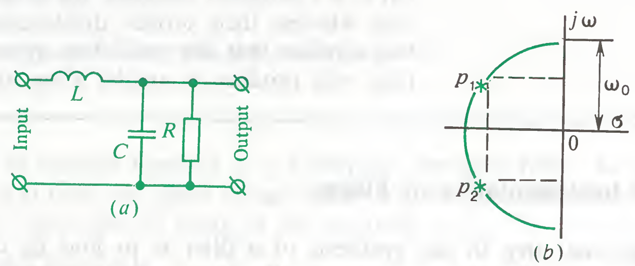


Fig. 13.7 A 2nd-order element: (*a*) schematic diagram; (*b*) location of the poles of the transfer function

Work Problem 10

It is an easy matter to find that for this type

$$K(p) = \frac{\omega_0^2}{p^2 + 2\alpha p + \omega_0^2} \tag{13.53}$$

where

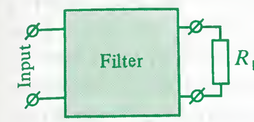
$\omega_0 = 1/\sqrt{LC}$ and $\alpha = 1/2RC$

The poles of the transfer function are

$$p_{1,2} = -\alpha \pm j\sqrt{\omega_0^2 - \alpha^2} \tag{13.54}$$

which may be complex conjugate or real, depending on the relative magnitude of ω_0 and α.

Let us consider several examples of a low-pass filter implemented with cascade-connected networks.



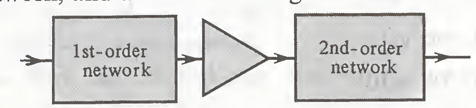
Example 13.10. *Realize an $n = 3$ Butterworth filter with a cut-off frequency of $10^5\ s^{-1}$. The filter is loaded into a resistor* $R_L = 0.5$ KΩ.

As has been shown, the transfer function of such a filter has three poles located as follows:

$$p_{1,2} = 10^5(\cos 60° \pm j\sin 60°)$$
$$= -0.5\times 10^5 \pm j0.866\times 10^5\ s^{-1}$$
$$p_3 = -10^5\ s^{-1}$$

Here, we have changed from the normalized complex variable p_N to the true complex frequency $p = \omega_c p_N$.

Accordingly, the sought-for filter may be visualized as a cascade connection of a 1st-order stage corresponding to the pole p_3, an isolation network, and a 2nd-order stage for which the poles are p_1 and p_2:



In accordance with (13.52), the 1st-order stage must have a time constant $RC = 1/\omega_c = 10^{-5}$ s. If we choose $C = 10$ nF, then the resistor used in this stage will be $R = 10^{-5}/C = 1$ kΩ.

Let the resistor used in the 2nd-order stage be the load resistor. On the basis of (13.54), the pair of complex conjugate roots will have the desired real part if

Work Problem 11

$$1/2R_L C = -\operatorname{Re} p_{1,2} = 0.5\times 10^5$$

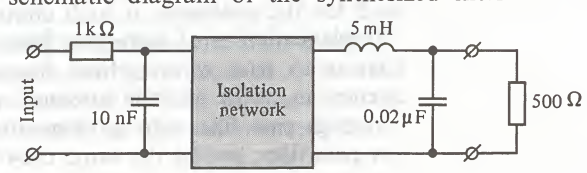
Hence,

$$C = 1/10^6 R_L = 0.02\ \mu F$$

Finally, the inductance is

$$L = 1/\omega_0^2 C = 5\ \text{mH}$$

The schematic diagram of the synthesized filter has the form



Example 13.11. *Realize an $n=2$ Chebyshev low-pass filter operating into a resistive load, $R_L = 1$ kΩ. The initial specifications for the synthesis are: cut-off frequency, $\omega_c = 10^5$ s^{-1}; ripple factor, $\varepsilon = 1$.*

In order to realize the second-order frequency response, it will suffice to have one L-section RLC-network. In Example 13.9 we have obtained the coordinates of the poles of the transfer function for an $n=2$ Chebyshev filter with $\varepsilon = 1$:

$$p_{N1,2} = -0.322 \pm j\,0.777$$

or, on changing back to the non-normalized variable p,

$$p_{1,2} = (-0.322 \pm j\,0.777) \times 10^5 \text{ s}^{-1}$$

▲ **Solve Problem 12**

The capacitance of the capacitor C can be found from (13.54) on equating α to the required abscissa of the poles:

$$\alpha = 1/2R_LC = 0.322 \times 10^5$$

Hence,

$$C = 15.53 \text{ nF}$$

The inductance L is found from the equation for the coordinates of the poles on the imaginary axis:

$$\sqrt{\omega_0^2 - \alpha^2} = 0.777 \times 10^5$$

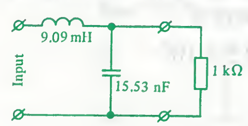
On solving it, we get

$$\omega_0^2 = 1/LC = 0.708 \times 10^{10}$$

$$L = 1/\omega_0^2 C = 9.09 \text{ mH}$$

▲ **Solve Problem 13**

Thus, the specified response characteristic is realized by the following circuit



It is to be noted that in practice, especially at microwave frequencies, use is made of filters in which there are no isolation networks. The reader may acquaint himself with the techniques used for the synthesis of such circuits in [29].

Implementation of high-pass filters. The function of a high-pass filter is to pass waves whose frequencies exceed the cut-off frequency, ω_c, with as little attenuation as practicable.

A high-pass filter can be generated directly from a normalized low-pass filter having the same cut-off frequency through a technique known in circuit theory as *frequency transformation*. The gist of the technique is as follows.

To begin with, we change from the variable p used in the synthesis of low-pass filters to a new frequency variable p' such that

$$p = \omega_c^2/p' \tag{13.55}$$

Frequency transformation ought not to be confused with the frequency conversion effected by nonlinear and parametric networks

Then the point $p=0$ will map into a point at infinity in the p'-plane, and the two points $p_{1,2} = \pm j\omega_c$ on the imaginary axis will map into two points $p'_{1,2} = \mp j\omega_c$, the latter differing from the former only in sign. Therefore, we may expect that the magnitude response of the high-pass filter generated from a low-pass filter by frequency transformation defined in (13.55) will be that of a high-pass filter.

Now each capacitor which has admittance pC in the low-pass filter must be replaced with an element whose admittance is $\omega_0^{-2}C/p'$, that is, with an inductor of inductance $L = 1/\omega_c^2 C$. Similarly, the inductor L in the low-pass filter must be replaced with a capacitor, $C = 1/\omega_c^2 L$. The resistive elements are left unchanged. The transformation just described is illustrated in Fig. 13.8.

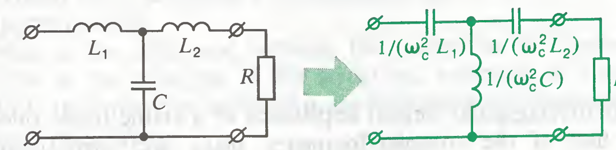


Fig. 13.8 Transformation of a low-pass filter into a high-pass filter

Implementation of band-pass filters. A band-pass filter passes with a small attenuation only the frequencies in the interval adjacent to some point $\omega_0 \neq 0$. If there is a normalized low-pass filter with the specified cut-off frequency, we can go over to a band-pass filter directly by a change of the variable

$$p = p' + \omega_0^2/p' \tag{13.56}$$

Frequency transformation entails some distortion in the amplitude response of the filter being synthesized, but it is of minor importance for a narrowband filter

Now the point $p=0$ will map into a point $p' = j\omega_0$, so that the peak of the magnitude response observed at zero frequency in the low-pass filter will occur at frequency ω_0 in the band-pass filter. Since

$$pC = p'C + \frac{\omega_0^2 C}{p'}$$

the admittance of the capacitor used in the low-pass filter must be replaced in the band-pass filter with the admittance of a parallel resonant circuit made up of a capacitor C and an inductor of inductance $L = 1/\omega_0^2 C$. It is to be noted that this resonant circuit is tuned to frequency ω_0.

Similarly from the equality

$$pL = p'L + \omega_0^2 L/p'$$

we conclude that the inductor L is to be changed into a series combination of the same inductor and a capacitor $C = 1/\omega_0^2 L$, which is recognized as a series resonant circuit tuned to frequency ω_0 (Fig. 13.9).

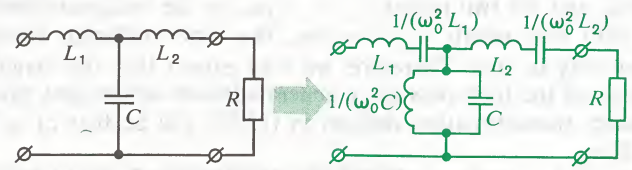


Fig. 13.9 Transformation from a low-pass filter into a band-pass filter

The prototype filter

From the examples examined above it follows that in the synthesis of frequency-selective networks the low-pass filter serves as the *prototype filter* whose parameters make it possible to generate subsequently any other form of filter.

Summary

- The poles and zeros of the driving-point (input) impedance of a stable linear one-port are located in the left half of the complex-frequency plane and form complex-conjugate pairs.
- The number of poles of the driving-point impedance of a passive one-port differs from the number of zeros by not more than one.
- The real and imaginary parts of the driving-point impedance of a one-port are related by a Hilbert transform pair, if the given one-port belongs to the class of minimum-impedance networks.
- The driving-point impedance of a reactive one-port is a non-decreasing function of frequency (Foster's theorem).
- The synthesis of a one-port from the specified driving-point impedance is effected within the framework of the specified circuit structure. This can be done by the Foster (or partial fraction expansion) method or by the Cauer (continued-fraction) method.
- The poles of the transfer function of a stable two-port (four-terminal) network are located only in the left half-plane.
- The zeros of the transfer function of a stable two-port may be located in the right half-plane (nonminimum-phase networks).
- The poles and zeros of the power transfer function show quadrant symmetry.
- The poles of the transfer function of a maximally-flat (Butterworth) low-pass filter are located on the unit circle whose radius is equal to the cut-off frequency.
- The poles of the transfer function of an equal-ripple (Chebyshev) filter are located on an ellipse whose eccentricity is given by the ripple factor of the magnitude response.
- High-pass and band-pass filters can be generated from a known low-pass filter which plays the role of a prototype.

Review Questions

1. Write the characteristic equations for a one-port which is (a) open-circuited and (b) short-circuited.
2. Define the property of a positive real (p.r.) function.
3. What one-ports are called minimum-impedance networks?
4. Name the basic properties of the poles and zeros of reactive one-ports.
5. Give an example of a ladder network and show how its driving-point immitance (impedance or admittance) can be written in the form of a continued fraction expansion.
6. What one-ports are called canonical?
7. Name the network functions used to characterize two-ports.
8. Why is it that the transfer function of a stable two-port must have a zero at infinity?
9. Why is it that any ladder two-port is a minimum-phase network?
10. What is the relation between the magnitude and phase responses of a minimum-phase two-port?
11. Name the engineering applications of nonminimum-phase networks.
12. How are the poles of the transfer function of a Butterworth low-pass filter located?
13. Explain why Chebyshev polynomials are convenient for use in the low-pass filter approximation.
14. What is the difference between the properties of Chebyshev and Butterworth filters?
15. What is the principle of the structural synthesis of filters?
16. Write the equations involved in the frequency transformation of low-pass to high-pass and band-pass filters.

Problems

1. Verify Eq. (13.16) with reference to the parallel RC-network examined in Example 13.1.

2. Show that the function

$$Z(p) = \frac{(p^2+1)(p^2+5)}{p(p^2+3)(p^2+7)}$$

is the driving-point (input) impedance of a realizable purely reactive one-port.

3. Implement the Foster-form network whose driving-point impedance is

$$Z(p) = \frac{p(p^2+4)}{p^2+1}$$

4. Implement two Cauer-form networks whose driving-point impedances are specified in Problem 3.

5. Prove that the first $n-1$ derivatives of the power transfer function of an nth-order Butterworth filter vanish at $\omega_N = 0$.

6. A Butterworth filter has a cut-off frequency of 10 kHz. In changing from 80 kHz to 160 kHz, the harmonic attenuation is increased by -36 dB. Find the order of the filter.

7. Find the transfer function of a 4th-order Butterworth filter.

8. A Chebyshev filter has a ripple factor $\varepsilon = 0.3$. What is the ripple factor of the same filter at $\omega_N < 1$ in decibels?

9. Find the transfer function of an $n = 2$ Chebyshev filter for $\varepsilon = 0.5$.

10. Show that a change in the resistance of the resistor R in a 2nd-order network causes the complex-conjugate poles of the transfer function to move round a circle of

radius ω_0. Analyse also the case of $\alpha > \omega_0$.

11. Implement a 2nd-order Butterworth filter with a cut-off frequency $\omega_c = 10^6\ s^{-1}$. The filter is loaded into a resistor $R_L = 20\ k\Omega$. When specifying the element values, make sure that the network is physically realizable.

12. Locate the poles of the transfer function of an $n = 3$ Chebyshev low-pass filter for which $\omega_c = 4 \times 10^4\ s^{-1}$ and $\varepsilon = 0.5$.

13. Implement the filter examined in Problem 12 by connecting in cascade a 1st-order and a 2nd-order network. The load resistance of the filter is $R_L = 2\ k\Omega$.

Advanced Problems

14. The real part of the driving-point impedance is

$$\frac{100}{1 + 10^{-12}\omega^2 + 10^{-24}\omega^4}$$

Find the expression for the complex driving-point impedance and draw up the corresponding circuit diagram.

15. Investigate the driving-point impedance $Z(p)$ of a semi-infinite ladder network with a periodic structure of elements.

16. There is a system which performs an ideal delay of signals for T seconds. Its transfer function is

$$K(p) = \exp(-pT) = \frac{1}{1 + pT + (pT)^2/2! + \ldots}$$

Check to see if this characteristic can approximately be replaced with a rational function whose denominator has various orders. Find the poles of the transfer function to a first, second, and third approximation.

17. Investigate the phase response of a Butterworth low-pass filter. Derive the equation for the group delay time for $\omega = 0$ and $\omega = \omega_c$.

18. Calculate the impulse responses for 2nd-order Butterworth and Chebyshev filters.

19. Derive an equation defining the asymptotic slope in decibels for the response of a Chebyshev filter at frequencies substantially exceeding the cut-off frequency.

Chapter 14

Active Networks with Feedback. Self-Excited Oscillatory Systems

This Chapter will be concerned with a special class of active linear and nonlinear circuits in which all or a part of the output signal is fed back to the input. Quite aptly, they are called *circuits* or *networks with feedback.*

The application of feedback makes it possible in some cases to improve the performance of the circuit involved to a marked degree. In other cases, given certain conditions, feedback may render the circuit unstable so that it jumps into self-excited oscillations. This principle underlies the operation of various self-excited oscillatory systems, above all harmonic oscillators which are integral parts of transmitters.

14.1 The Transfer Function of a Linear Feedback System

In order to make the results of the analysis that follows applicable to a wide range of various special cases, we will consider the problem of a feedback circuit in the most general terms, without exactly specifying the physical character of the input signal (excitation) and the output signal (response).

The derivation of the basic relation. We will be concerned with the linear system shown in block-diagram form in Fig. 14.1.

The forward-path and the feedback elements

Referring to the diagram, the system is seen to consist of two four-terminal networks. The active two-port whose transfer function is $K(p)$ is the *forward-path element* of the system. The other two-port, usually a passive one, is called the *feedback element* with a transfer function (the feedback ratio) $\beta(p)$. The arrowheads in the figure indicate the direction of signal flow around the system.

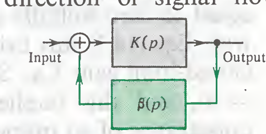


Fig. 14.1 Block diagram of a linear feedback system

At its input, the forward-path element has a network which combines the input signal and the output signal from the feedback element. If $V_{in}(p)$ and $V_{out}(p)$ are the Laplace transforms of the input and output signals, respectively, it is easy to see that

$$V_{out}(p) = K(p)\,[V_{in}(p) + \beta(p)\,V_{out}(p)] \tag{14.1}$$

Hence we can directly derive the equation defining the transfer function of a system with feedback:

$$K_{\text{fb}}(p) = V_{\text{out}}(p)/V_{\text{in}}(p) = \frac{K(p)}{1 - \beta(p)K(p)} \tag{14.2}$$

In accordance with the above equation, the frequency behaviour of the system depends equally on the transfer function $K(p)$ and the feedback ratio $\beta(p)$. Therefore, it is possible, while leaving the forward-path element unchanged, to vary the frequency characteristics of the entire system over a wide range by varying the parameters of only the feedback element.

Negative and positive feedback. Examine Eq. (14.2) for $p = j\omega$. The frequency response of a system with feedback* is

$$K_{\text{fb}}(j\omega) = \frac{K(j\omega)}{1 - \beta(j\omega)K(j\omega)} \tag{14.3}$$

If at a specified frequency ω,

Solve Problem 1

$$|1 - \beta(j\omega)K(j\omega)| > 1 \tag{14.4}$$

then feedback will reduce the magnitude response of the system and, as a consequence, the amplitude of the output signal. This is known as *negative (degenerative) feedback*. If the converse is true, that is

$$|1 - \beta(j\omega)K(j\omega)| < 1 \tag{14.5}$$

we have *positive (regenerative) feedback*.

Sometimes, instead of the terms "positive" and "negative" feedback, the wider concept of "complex" feedback is used. In this case, the value of the phase shift in the feedback element is specified

Both forms of feedback are widely used in communication applications. It should be borne in mind, however, that positive feedback may cause instability in the system. To demonstrate, let, for example, $\beta = \beta_0$ and $K = K_0$ be positive real numbers. If β_0 is equal to zero initially and increases afterwards, then in accordance with Eq. (14.3) this brings about an increase in the overall gain or closed-loop gain K_{fb}. Should β_0 become equal to $1/K_0$, then $K_{\text{fb}} = \infty$, and this implies the self-excitation of the system, or the appearance of an output signal when there is no signal at the input.

Negative feedback makes it possible to improve considerably the frequency response of amplifiers. The examples that follow will

* In engineering applications, this quantity is usually referred to as the *closed-loop gain* or the *overall gain with feedback.–Translator's note.*

demonstrate some of the specific applications where negative feedback is warranted.

Gain stabilization. Suppose we have an amplifier with a large, but insufficiently stable gain K_0. It is required to build on its basis an amplifier with an improved gain stability. By applying a negative feedback loop to the amplifier, that is, by taking $\beta(j\omega) = -\beta_0 < 0$, we obtain on the basis of (14.3)

$$K_{\text{fb}} = K_0/(1 + \beta_0 K_0)$$

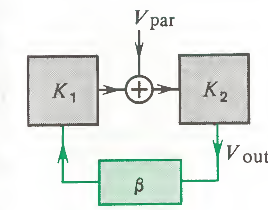
Hence,

$$dK_{\text{fb}}/K_{\text{fb}} = [1/(1 + \beta_0 K_0)]\,(dK_0/K_0) \tag{14.6}$$

Solve Problem 2

If $\beta_0 K_0 \gg 1$, the relative instability in the resultant gain decreases by a factor of about $\beta_0 K_0$. True, the gain itself decreases by the same factor, but this does not usually pose any problem because the desired gain can always be secured by adding more stages to the amplifier.

Suppression of parasitic signals. Let the forward-path element of an amplifier be a cascade connection of two networks of gains K_1 and K_2, respectively. At their junction, an unwanted parasitic signal V_{par} is injected. A negative feedback loop with the fraction fed back β is applied to the amplifier as a whole. It is required to determine the parasitic-signal gain $K_{\text{par}} = V_{\text{out}}/V_{\text{par}}$, that is the extent to which the parasitic signal is allowed to reach output.



Since obviously

$$V_{\text{out}} = K_2(V_{\text{par}} - \beta K_1 V_{\text{out}})$$

then

$$K_{\text{par}} = V_{\text{out}}/V_{\text{par}} = K_2/(1 + \beta K_1 K_2) \tag{14.7}$$

It is seen that the parasitic signal breaking through into the system at a point close to its output, that is, with $K_2 \ll K_1$, will be substantially suppressed. This effect provides the basis for the control of nonlinear distortion in multistage amplifiers (Fig. 14.2).

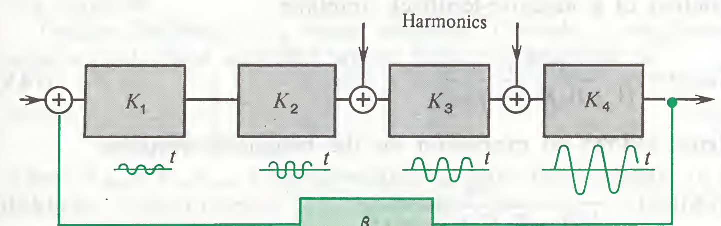


Fig. 14.2 Harmonic suppression in a multistage negative-feedback amplifier

As will be recalled (see Chap. 11), nonlinear distortion manifests itself as the appearance of signal harmonics because of the nonlinearity of the active elements. The harmonic content goes up as the signal amplitude is increased. Mentally, it may be pictured as if the unwanted harmonics are injected into the system from without, mainly in the final power stages of the amplifier. On the basis of (14.7) we conclude that negative feedback can substantially reduce the harmonic content at the output. Therefore, practically all amplifiers intended for the high-fidelity reproduction of audio signals (in radio broadcasting, sound recording) use negative feedback.

In high-fidelity (hi-fi) amplifiers the non-linear distortion is a fraction of one per cent

Frequency response compensation. Consider a single-stage RC-loaded transistor amplifier whose transfer function (see Eq. (8.43)) is

$$K(p) = -K_0/(1 + p\tau_{eq}) \tag{14.8}$$

where $K_0 = g_m R_{eq}$, and $\tau_{eq} = R_{eq} C_s$.

At zero frequency, the gain factor is negative

$$K(0) = -K_0$$

By applying to this amplifier a frequency-independent feedback loop with a positive real parameter $\beta(j\omega) = \beta_0$, we will, on the basis of (14.2), have

$$K_{fb}(0) = -K_0/(1 + \beta_0 K_0)$$

Since $1 + \beta_0 K_0 > 1$, we have negative feedback. It is easy to verify that the feedback will remain negative at all frequencies because

Here, negative feedback exists, owing to the sign of the gain at the zero frequency

$$|1 - \beta_0 K(j\omega)| = \left|1 + \frac{\beta_0 K_0}{1 + j\omega\tau_{eq}}\right| > 1$$

at any frequency $\omega > 0$. On substituting (14.8) in the general equation (14.2), we obtain the following expression for the transfer function of a negative-feedback amplifier:

$$K_{fb}(p) = \frac{-K_0}{(1 + \beta_0 K_0) + p\tau_{eq}} \tag{14.9}$$

Hence follows an expression for the magnitude response

$$|K_{fb}(j\omega)| = \frac{K_0}{\sqrt{(1 + \beta_0 K_0)^2 + \omega^2\tau_{eq}^2}} \tag{14.10}$$

Figure 14.3 shows a set of amplitude response curves for amplifiers differing in the negative feedback level as defined by the

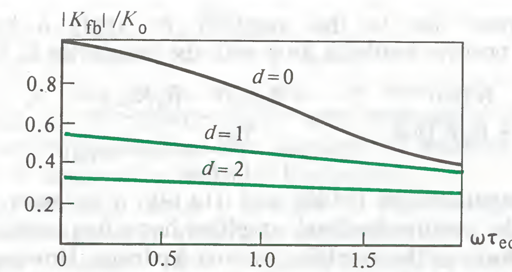


Fig. 14.3 Amplitude response of a single-stage RC-loaded amplifier for various levels of negative feedback

parameter*

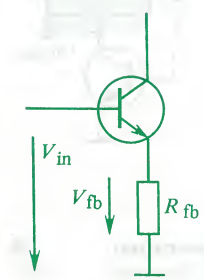
$$d = \beta_0 K_0 \tag{14.11}$$

The figure brings out the principal effect of negative feedback: the amplitude response of the amplifier is "equalized", or compensated, owing to a reduction in gain at lower frequencies. As a consequence, the effective bandwidth of the amplifier is extended. Thus, on the basis of (14.10) the upper limiting, or cut-off, frequency ω_c, defined as one at which the gain of the amplifier falls to 0.707 times the maximum gain, K_0, is

$$\omega_c = (1 + \beta_0 K_0)/\tau_{eq} \tag{14.12}$$

and increases linearly with increasing negative-feedback factor.

The simplest way to apply negative feedback to a single-stage common-emitter amplifier is to place an additional feedback resistor R_{fb} in the emitter lead. The resultant increase in the input resistance brings about an increase in the emitter current and, in consequence, an increase in V_{fb}, the voltage across the feedback resistor. Therefore, the control voltage of the transistor becomes



$$V_{BE} = V_{in} - V_{fb}$$

which is another way of saying that negative feedback does exist in the amplifier.

▲ Solve Problem 3

Positive feedback in a tuned amplifier. Consider a single-stage small-signal tuned amplifier whose frequency response is

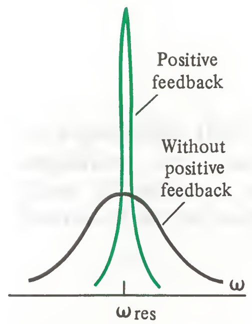
$$K(j\omega) = \frac{-K_{res}}{1 + j\tau_{ckt}(\omega - \omega_{res})} \tag{14.13}$$

Only positive frequencies are considered

where $K_{res} = g_m R_{res}$, and $\tau_{ckt} = 2Q/\omega_{res}$ is the time constant of the resonant (tuned) circuit.

---

* In books on practical radio engineering, it is often called the *feedback factor*, or *return ratio*, or *loop gain*.– *Translator's note.*



The effect of positive feedback on the amplitude response of an amplifier

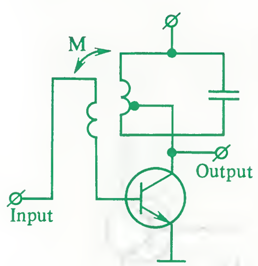
Now suppose that to this amplifier we apply a frequency-independent positive feedback loop with the parameter β_0 such that

$$K_{fb}(j\omega) = \frac{K(j\omega)}{1+\beta_0 K(j\omega)} = \frac{-K_{res}/(1-\beta_0 K_{res})}{1+j\dfrac{\tau_{ckt}}{1-\beta_0 K_{res}}(\omega-\omega_{res})} \tag{14.14}$$

From comparison of (14.13) and (14.14), it is seen that for $\beta_0 K_{res} < 1$, the positive-feedback amplifier has a frequency response of the same form as the amplifier without feedback. However, in the case of positive feedback there is an increase in the gain at resonance by a factor $1/(1-\beta_0 K_{res})$; the equivalent Q-factor of the resonant circuit of the amplifier is increased in the same proportion

$$Q_{eq\,fb} = Q_{eq}/(1-\beta_0 K_{res})$$

whereas the bandwidth of the amplifier is reduced by the same factor.



Regeneration

These changes are explained by the fact that positive feedback produces regeneration – a partial compensation of the losses suffered by the resonant circuit. The energy required for the regeneration is drawn from the power supply.

Positive feedback in a tuned amplifier can be produced by placing an inductor connected in series with the input circuit and inductively coupled with the resonant circuit.

Despite a number of obvious advantages, positive-feedback amplifiers are used only seldom because they tend to jump into self-excited oscillation as the feedback factor $\beta_0 K_{res}$ goes to unity.

Delayed feedback. Figure 14.4 shows a block diagram of a system in which the feedback path contains a scaling amplifier with a constant fraction fed back, β_0, and a pure delay element which delays signals for a time τ_0.

Let the gain K_0 of the forward-path element be frequency independent. Then

$$K_{fb}(j\omega) = \frac{K_0}{1-\beta_0 K_0 \exp(-j\omega\tau_0)} \tag{14.15}$$

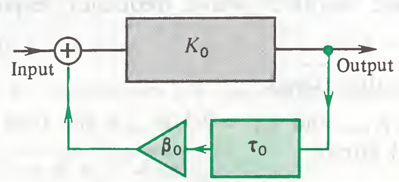
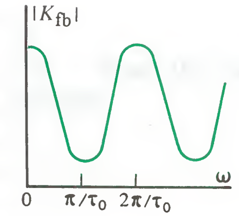


Fig. 14.4 Block diagram of a delayed-feedback system



The amplitude response of the comb filter

The amplitude response of the system is

$$|K_{fb}(j\omega)| = \frac{K_0}{\sqrt{1-\beta_0 K_0 \cos\omega\tau_0 + (\beta_0 K_0)^2}} \tag{14.16}$$

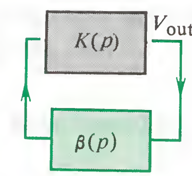
If $\beta_0 K_0 < 1$, the system is stable. Its amplitude response is represented by a periodic curve with alternating maxima and minima which implies that the character of feedback is different (positive or negative) at different frequencies.

Delayed feedback makes it possible to build frequency-selective systems with a periodic amplitude response, known as *comb filters* so named because their response characteristics have the appearance of a comb.

It is to be noted that this type of systems are capable of jumping into self-excited oscillation as $\beta_0 K_0 \to 1$.

14.2 Stability of Feedback Networks

This section will discuss the stability of feedback systems. The objective of the discussion is to confirm the qualitative reasoning pursued earlier with regard to self-excitation of positive feedback systems.



The statement of the problem. Consider a system made up of an active forward-path element whose transfer function is $K(p)$ and a feedback element whose transfer function is $\beta(p)$. The output of the forward-path element is returned to the system's input via the feedback element. It is assumed that no external input signal is applied, so that the system is self-contained.

The equation of state is written on the basis of the fact that

$$V_{out}(p) = K(p)\beta(p)V_{out}(p)$$

Hence,

$$[1-\beta(p)K(p)]\,V_{out}(p) = 0 \tag{14.17}$$

The characteristic equation of a feedback system

Since $V_{out}(p) \neq 0$ identically (otherwise the system would not be excited), then the equality (14.17) will be satisfied only for those values of p which are the roots of the characteristic equation

$$1-\beta(p)K(p) = 0 \tag{14.18}$$

Let $p_1, p_2, \ldots$ be the roots of the characteristic equation. Since the system is linear, the output signal will in the general case have the form

$$v_{out}(t) = A_1\exp(p_1 t) + A_2\exp(p_2 t) + \ldots \tag{14.19}$$

For the signal to be bounded, it is necessary that all the roots of the characteristic equation have negative real parts, that is, be located in the left half p-plane. A feedback system possessing such

properties will be absolutely stable in the sense defined in Chap. 8.

Two problems may arise in the study of feedback systems. If the system being synthesized, say an amplifier, must be stable, we need a criterion with which we could immediately tell from the form of the functions $\beta(p)$ and $K(p)$ that the characteristic equation has no roots lying in the right half p-plane. Conversely, if feedback is utilized to build an unstable self-excited oscillatory system, we need to know the roots of Eq. (14.18) defining the frequency at which the system will jump into self-oscillation.

This section will be concerned with the first of the two problems. It is to be noted that the conclusions derived here apply not only to the stability of feedback systems, but to the stability of any linear dynamic system as well.

Algebraic criteria of stability. Suppose that both the forward-path and the feedback elements are lumped-parameter networks and so

$$K(p) = P_1(p)/Q_1(p)$$

$$\beta(p) = P_2(p)/Q_2(p) \tag{14.20}$$

are the ratios of the polynomials in powers of p. On substituting (14.20) into (14.18), we obtain the characteristic equation of the system

$$\frac{Q_1(p)Q_2(p) - P_1(p)P_2(p)}{Q_1(p)Q_2(p)} = 0 \tag{14.21}$$

Hence it follows that a feedback system is stable if all roots of the equation

$$H(p) = Q_1(p)Q_2(p) - P_1(p)P_2(p) = 0$$

have negative real parts. In algebra, polynomials $H(p)$ with such properties are called *Hurwitz polynomials*.

Hurwitz polynomials

Consider a special case of the Hurwitz polynomial

$$H(p) = (p - p_1)(p - p_2)(p - p_3)$$

which has three roots one of which, $p_1 = -\alpha$, is real negative, whereas the remaining two, $p_{2,3} = -\beta \pm j\omega_0$, are complex conjugate numbers with a negative real part. On direct substitution of the roots, we find that the polynomial

$$\begin{aligned} H(p) &= (p + \alpha)[(p + \beta)^2 + \omega_0^2] \\ &= p^3 + (\alpha + 2\beta)p^2 + (\beta^2 + 2\alpha\beta + \omega_0^2)p + \alpha(\beta^2 + \omega_0^2) \end{aligned}$$

contains all powers of the variable p, beginning with the highest one, and all of its coefficients have the same sign. These attributes only indicate the necessary conditions for a polynomial to be a Hurwitz polynomial. The complete solution of the problem was

obtained at the end of the last century and has been embodied in what is known as the *Routh-Hurwitz* criterion. The proof of the criterion can be found in [11]. We will only give the final result: For the equation

The Routh-Hurwitz criterion

$$a_n p^n + a_{n-1}p^{n-1} + \ldots + a_1 p + a_0 = 0$$

with real coefficients to have roots lying in the left half p-plane, it is necessary and sufficient that the following quantities be positive:

(1) The coefficients a_n, a_0.

(2) The Routh-Hurwitz determinant

$$D_{n-1} = \begin{vmatrix} a_{n-1} & a_n & 0 & 0 & \ldots & 0 & 0 \\ a_{n-3} & a_{n-2} & a_{n-1} & a_n & \ldots & 0 & 0 \\ a_{n-5} & a_{n-4} & a_{n-3} & a_{n-2} & \ldots & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \ldots\ldots & & a_0 & a_1 \end{vmatrix}$$

and all of its principal minors.

Example 14.1. *Using the Routh-Hurwitz criterion, test the stability of a system whose characteristic equation has the form*

$$p^3 + 2p^2 + 6p + 4 = 0$$

We satisfy ourselves that a_3, $a_0 > 0$. Next, we form the determinant

$$D_2 = \begin{vmatrix} 2 & 1 \\ 4 & 6 \end{vmatrix} = 8 > 0$$

Solve Problem 6

The only principal minor is $2 > 0$. Hence the system is stable.

An advantage of the Routh-Hurwitz criterion is the relative simplicity of calculations. A disadvantage is that its applicability is limited to lumped-constant networks because it is only for them that the transfer function is expressed in terms of polynomials.

Graphical criteria of stability. Going back to the characteristic equation (14.18) we note that the product

$$w(p) = \beta(p)K(p) \tag{14.22}$$

is nothing but the transfer function of a cascaded connection of a forward-path element and a feedback element. Usually, $w(p)$ is called the open-loop transfer function of a feedback system.

The function in (14.22) may be looked upon as a mapping of the complex p-plane onto another complex w-plane. If $p_1, p_2, \ldots$ are the roots of the characteristic equation

$$1 - \beta(p)K(p) = 0$$

then, as can be readily seen, all of these points will be mapped in the w-plane into an only point, $w = 1$.

Hence we can directly formulate the rule for determining whether a feedback system is capable of self-excitation: If, under the conformal mapping of (14.22), the w-plane image of the right half p-plane contains the point $w = 1$, the feedback system is closed-loop unstable.

The imaginary ($j\omega$) axis in the p-plane maps into a curve in the w-plane. The equation of the curve in parametric form is

$$w(j\omega) = \beta(j\omega)K(j\omega) \quad (14.23)$$

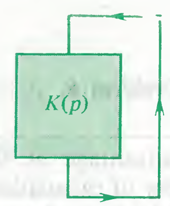
The Nyquist locus

The parameter is frequency ω which varies between the limits $-\infty$ and $+\infty$. The curve is called the *Nyquist locus* of an open-loop feedback system. In all cases of practical interest, the amplitude response tends to zero with increasing frequency, so the Nyquist locus passes through the point $w = 0$. Also, the Nyquist locus is symmetrical about the real axis in the w-plane, because $w(-j\omega) = w^*(j\omega)$. Obviously, the Nyquist loci for the systems in question are closed curves in the w-plane.

The Nyquist stability criterion

In the theory of functions of a complex variable it is shown [11] that under the conformal mapping of (14.22) the image of the right half-plane is the area enclosed by the Nyquist locus. The stability criterion arising from the above procedure is known as the *Nyquist stability criterion*: If the Nyquist locus of an open-loop system encloses the point (0, 1), the system is closed-loop unstable.

Example 14.2. *Investigate for stability a single-stage RC-coupled amplifier whose output is directly connected to its input.*



Here, obviously, $\beta(p) = 1$, and

$$K(p) = -K_0/(1 + p\tau)$$

where

$K_0 = g_m R_{eq}$ and $\tau = R_{eq}C_s$ (see Chap. 8)

The equation of the Nyquist locus takes the form

$$w(j\omega) = \frac{K_0}{\sqrt{1 + \omega^2\tau^2}} \exp[j(\pi - \arctan\omega\tau)] \quad (14.24)$$

The Nyquist locus plotted on the basis of Eq. (14.24) appears in Fig. 14.5. As is seen, the locus diagram is a circle of diameter K_0. The top semi-circle is the image of the positive part of the $j\omega$-axis. As the frequency rises, the amplitude response decreases, and the phase angle tends to 90°.

Since the Nyquist locus is entirely in the left half-plane and does not enclose the $w = 1$ point, the connection of the output of the amplifier to its input leaves the system stable.

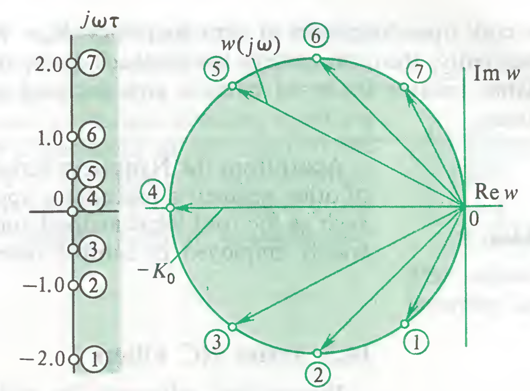
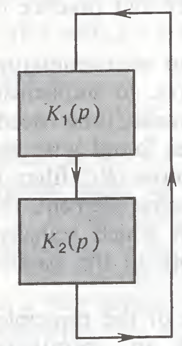


Fig. 14.5 Nyquist locus diagram of a single-stage RC-loaded amplifier. (The numerals in circles indicate corresponding points on the $j\omega$-axis and on the Nyquist locus.)



Solve Problem 7

Example 14.3. *Using the Nyquist stability criterion, investigate for stability a two-stage aperiodically loaded amplifier whose transfer function is*

$$K(p) = K_1(p)K_2(p) = K_{01}K_{02}/(1 + p\tau)^2 \quad (14.25)$$

To simplify the matters, it is assumed that the two stages have the same time constant; the gain constants at zero frequency, K_{01} and K_{02}, may in the general case be different.

By a change of variable $p = j\omega$ in (14.25), we obtain the equation for the Nyquist locus

$$w(j\omega) = K_{01}K_{02}/(1 + j\omega\tau)^2 = \frac{K_{01}K_{02}}{1 + \omega^2\tau^2}\exp(-j2\arctan\omega\tau) \quad (14.26)$$

The Nyquist locus diagram for the amplifier in question appears in Fig. 14.6.

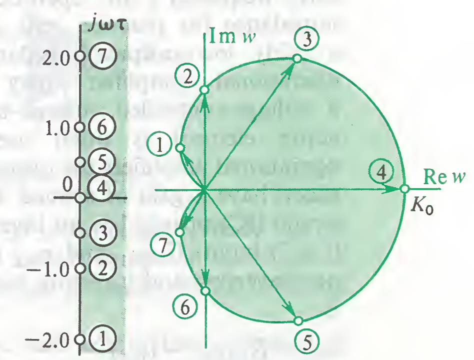


Fig. 14.6 The Nyquist plot of a two-stage, aperiodically loaded amplifier

If the overall open-loop gain at zero frequency $K_0 = K_{01}K_{02}$ is greater than unity, then, on closing the feedback loop, the system turns unstable, because the $w = 1$ point is now enclosed within the Nyquist locus.

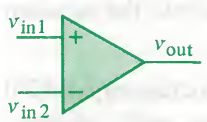
▲ Solve Problem 8

Apart from the Nyquist criterion, stability analysis uses a number of other geometric procedures applicable to linear feedback systems, such as the root-locus method and the Bode method [37]. They are widely employed in control system analysis.

14.3 Active *RC* Filters

Present-day advances in radio engineering and electronics, notably the advent of microelectronic components, have brought about sweeping changes in the previously widely used circuit designs. The changes have also affected the theory and practice of frequency-selective filter synthesis.

Among other things, it has been found that no microminiature counterpart of the inductor can be fabricated. Yet, to implement conventional 2nd-order oscillatory (*RLC*) networks, one needs inductances (see Chap. 13). The way out has been found with the development of so-called active *RC* filters. An active *RC* filter is a combination of a passive *RC*-network and an active element, the latter usually being a complex transistor circuit which transfers some of the power it draws from a power supply to the passive network.



The opamp

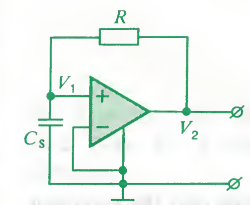
In this section we will be concerned with one of the principles used to build active *RC* filters which include an operational amplifier used as the active element.

The operational amplifier. This term refers to an amplifier with a large gain K_0 over a broad interval of frequencies, beginning at zero frequency. An operational amplifier has a high input impedance (in practice, tens or even hundreds of kilohms) and a fairly low output impedance (tens of ohms). Therefore, an operational amplifier may approximately be treated as a voltage-controlled voltage source. This model of a controlled active element is often used in circuit theory. Present-day operational amplifiers (or *opamps*, as they are frequently termed for short) have a gain of around 10^4 or 10^5. As a rule, an integrated-circuit (IC) opamp has an inverting ("–") input and a non-inverting ("+") input. If v_{in1} and v_{in2} are the input signal voltages at the non-inverting and inverting inputs, respectively, the output voltage is

$$v_{out} = K_0(v_{in1} - v_{in2}) \tag{14.27}$$

Opamps are among the most widely used analog IC's.

Stability of a system with an opamp. If feedback is applied around a system incorporating an opamp, a condition of instability is very likely to develop. As an example, consider a system in which the output of the opamp is connected to the non-inverting input via a resistor R. Let C_s be the stray capacitance at the input of the opamp. Since the input impedance of the opamp is infinitely high, the Laplace transform of the input voltage, $V_1(p)$, is related to the Laplace transform of the output voltage, $V_2(p)$, in a simple way



The signal at the inverting input is zero

$$V_1 = V_2\frac{1/pC_s}{1/pC_s + R}$$

On the other hand,

$$V_1 = V_2/K_0$$

The two equalities can be satisfied at the same time only if p is the root of the characteristic equation

$$1/(1 + pRC_s) = 1/K_0$$

that is,

$$p = (K_0 - 1)/RC_s$$

Thus, at $K_0 > 1$, the system is unstable; the voltages around the system build up in time as $\exp[(K_0 - 1)t/RC_s]$.

It is clear that a system in which the output of the opamp is connected to the inverting input via a resistor is stable.

The principle of synthesis of active RC networks. Let us examine in a very general way one of the procedures for the synthesis of an active filter on the basis of an opamp (Fig. 14.7). This procedure is applicable to a large number of practical circuit designs.

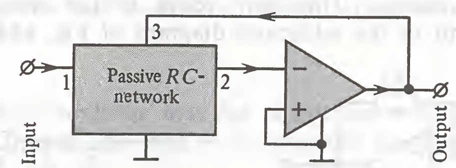


Fig. 14.7 Active *RC*-network built around an opamp

The passive part of the device is a three-port made up of elements R and C. Terminal *1* is the input; the opamp is connected between terminals *2* and *3*, so that use is made of the inverting input.

This circuit is a special case of the feedback system examined at the beginning of this chapter. In order to find the transfer function of the system in question, we will describe the passive three-port

with the aid of its Y-matrix:

$$I_1 = Y_{11}V_1 + Y_{12}V_2 + Y_{13}V_3$$
$$I_2 = Y_{21}V_1 + Y_{22}V_2 + Y_{23}V_3 \qquad (14.28)$$
$$I_3 = Y_{31}V_1 + Y_{32}V_2 + Y_{33}V_3$$

Note that the inverting input of the opamp is used

If K_0 is the gain of the opamp, then

$$V_3 = -K_0 V_2$$

The input circuit of the opamp draws no current, so, on the basis of the second line in (14.28), we may write

$$0 = Y_{21}V_1 + (Y_{23} - Y_{22}/K_0)V_3$$

Hence, the transfer function of the system is

$$K(p) = V_3/V_1 = -Y_{21}/[Y_{23} - Y_{22}/K_0]$$

Assuming that $K_0 \gg 1$, we finally have

$$K(p) = -Y_{21}/Y_{23} \qquad (14.29)$$

Thus, the transfer function of an active RC filter is solely dependent on the properties of its passive network; the gain of the opamp and its other parameters are excluded from the final result. Therefore, the construction of systems varying in frequency response reduces to the synthesis of passive RC-networks (in our case, a three-port) having the specified frequency response. The synthesis of three-ports is not taken up in this book. Therefore, we will take a simpler path – we will investigate some specific systems arranged in simple circuit configurations, which illustrate the circuit-engineering potentialities of the respective class of networks.

The scale changer. This term refers to the circuit shown in simplified form in the schematic diagram of Fig. 14.8.

▲ Solve Problem 9

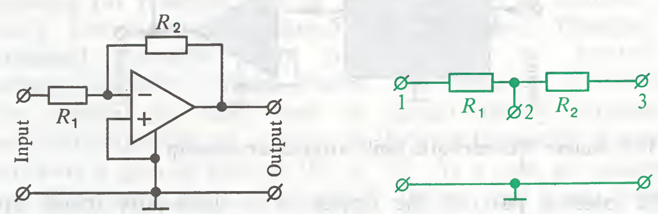


Fig. 14.8 Scale changer: (a) schematic diagram; (b) passive three-port

In order to determine the transfer admittances Y_{21} and Y_{23}, let us turn to the second line in Eqs. (14.28) and note that, say, $Y_{21} =$

$= I_2/V_1$ when terminals 2 and 3 are shorted to ground, that is, for $V_2 = V_3 = 0$. As is seen from Fig. 14.8b, $Y_{21} = 1/R_1$. Similarly, $Y_{23} = 1/R_2$. On substituting the two expressions in (14.29), we get

$$K(p) = -R_2/R_1 \qquad (14.30)$$

If $R_1 = R_2$, the device acts as a sign inverter

Equation (14.30) explains the term "scale changer": As is seen, by suitably varying the ratio of the resistors R_1 and R_2, the scale of gain (amplification) can be changed at will. Naturally, since the circuit does not contain any reactive elements, it has no frequency-selective properties.

The analog integrator. If, in the circuit examined above, we replace the resistor R_2 with a capacitor of capacitance C, we will obtain a circuit which performs the operation of electric integration on the input signal (Fig. 14.9). To demonstrate, from analysis of the passive network we find that $Y_{21} = 1/R$ and $Y_{23} = pC$. Hence, on the basis of Eq. (14.29), we have

$$K(p) = -1/pRC \qquad (14.31)$$

▲ Solve Problem 10

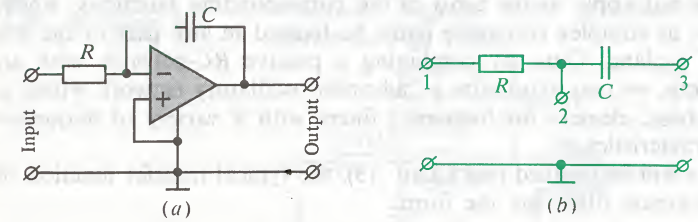


Fig. 14.9 Integrator built around an opamp: (a) schematic diagram; (b) passive three-port

Equation (14.31) confirms that the circuit has a transfer function which is the inverse of p, and so it integrates the signal applied to its input.

The low-pass filter. So that the frequency characteristics of active RC filters can be varied at will, it is necessary to use passive networks with a greater number of elements than in the above cases. An attractive example of a system which has the properties of a low-pass filter is shown in Fig. 14.10. An elementary analysis, carried out along the same lines as for the above cases, leads us to the following expression for the transfer function of the system

$$K(p) = \frac{-Y_1Y_3}{Y_5(Y_1 + Y_2 + Y_3 + Y_4) + Y_3Y_4} \qquad (14.32)$$

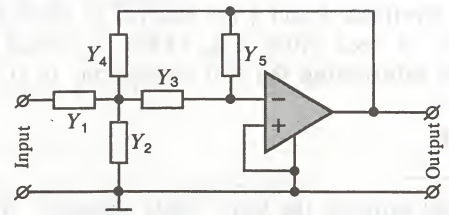


Fig. 14.10 Active *RC* low-pass filter

The difference between the poles and zeros of the functions of RC-networks

Thus, the task of synthesis reduces to matching the admittances of the elements in such a way that the desired frequency characteristic is implemented. From the initial specifications, all admittances Y_1 through Y_5 are either resistors of conductance G or capacitors whose admittances are pC. What is important in principle is that the transfer function of an active filter is expressed, in accordance with Eq. (14.29), by the ratio of two transfer admittances Y_{21} and Y_{23}. Then the poles of $K(p)$ coincide with the zeros of $Y_{23}(p)$. It is shown in circuit theory [27] that the poles of any driving-point or transfer function of a passive *RC*-network can be located only on the negative real axis. However this restriction does not apply *to the zeros* of the corresponding functions, which may, as complex conjugate pairs, be located in any part of the left half *p*-plane. Thus, by combining a passive *RC*-network with an opamp, we may synthesize a 2nd-order oscillatory network which is the basic element for frequency filters with a variety of frequency characteristics.

As will be recalled (see Chap. 13), the typical transfer function of a low-pass filter has the form

$$K(p) = \frac{A_0}{ap^2 + bp + c}$$

where A_0, a, b, and c are constants.

Referring to Eq. (14.32), it is seen that for the transfer function to have the above form, Y_1, Y_3 and Y_4 must be resistors, and Y_2 and Y_5 must be capacitors. Then,

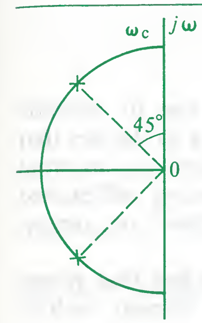
$$K(p) = \frac{-G_1G_3}{p^2C_2C_5 + pC_5(G_1 + G_3 + G_4) + G_3G_4} \tag{14.33}$$

Since the real parts of the poles are negative with any choice of parameters, the network is absolutely stable

The poles of the transfer function are located at points

$$p_{1,2} = -\frac{G_1 + G_3 + G_4}{2C_2} \pm j\sqrt{\frac{G_3G_4}{C_2C_5} - \frac{1}{4}\left(\frac{G_1 + G_3 + G_4}{C_2}\right)^2} \tag{14.34}$$

The above formula makes it possible to synthesize oscillatory elements with any predetermined location of poles.



Example 14.4. *Synthesize a second-order Butterworth active RC low-pass filter with a cut-off frequency* $\omega_c = 10^3\ \mathrm{s}^{-1}$.

As is stated in Chap. 13, the transfer function of such a filter should have two poles

$$p_{1,2} = \omega_c(-0.707 \pm j\,0.707) \tag{14.35}$$

Let us assume reasonable values for the resistors in the circuit and make them all equal, that is $R_1 = R_3 = R_4 = 1.8\ \mathrm{k\Omega}$. Then $G_1 = G_3 = G_4 = 5.55 \times 10^{-4}$ S.

On equating the real parts of (14.34) and (14.35), we obtain the expression defining the capacitance C_2

$$\frac{G_1 + G_3 + G_4}{2C_2} = 0.707\omega_c$$

Hence, on substituting the known values, we find that $C_2 = 1.18\ \mu\mathrm{F}$.

In order to determine the capacitance of capacitor C_5, we equate the imaginary parts of (14.34) and (14.35):

$$\frac{G_3G_4}{C_2C_5} - \frac{1}{4}\left(\frac{G_1 + G_3 + G_4}{C_2}\right)^2 = \omega_c^2/2$$

Solve Problem 11

On solving the above equation, we find that $C_5 = 0.26\ \mu\mathrm{F}$. The schematic diagram of the synthesized active Butterworth filter is shown in Fig. 14.11.

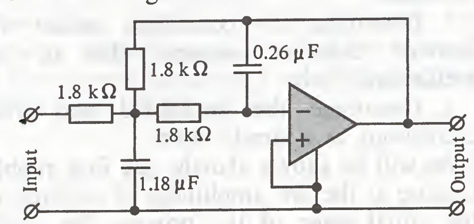


Fig. 14.11 Active *RC*-filter with a maximally flat response at a cut-off frequency of $10^3\ \mathrm{s}^{-1}$

Concluding remarks. In this section we have examined the synthesis of active *RC* filters on the basis of opamps, amplifiers with an infinitely high gain. However, this principle does not exhaust the capabilities of present-day integrated circuits. Of particular interest in this respect are *gyrators*, active two-ports which have the property that when a capacitor is connected to the output port, a purely inductive input impedance is synthesized. In this way, filters can be built which do not contain any physical inductive elements.

The gyrator

For the properties of gyrators and other active elements used in present-day *RC* filters the reader is referred to [32, 33].

14.4 Self-Excited Harmonic Oscillators. The Sma ll-Signal Condition

In some situations active networks can give rise to periodic *self-excited oscillations*. The term "self-excited" refers to the fact that the oscillations are produced and maintained without any external periodic excitation of the system. Devices generating self-excited oscillations are frequently called *self-excited oscillators* or, simply, *oscillators*.

For its operation any oscillator depends on the fact that power from an external power supply is fed via an active element, such as a transistor, to an oscillatory system (a resonant or tank circuit). The signal that controls the transistor is picked off the same resonant circuit and applied to the control electrode of the transistor over a feedback path.

With suitably chosen parameters, such a system becomes unstable. Any minute oscillations, such as caused by thermal noise, tend to grow in amplitude without bound. However, as their amplitude builds up, the nonlinear properties of the control element begin to play a progressively greater role, so that the amplitude reaches a steady-state value and remains practically constant afterwards. The oscillator is then said to operate in a *steady state*.

The steady state

The analysis and synthesis of oscillators poses two basic problems:

1. Determine the conditions under which the circuit with feedback becomes unstable, that is, jumps into self-excited oscillations.

2. Determine the amplitude and frequency of self-excited oscillations in a steady state.

As will be shown shortly, the first problem is simpler to solve, because at the low amplitudes of self-excited oscillations, typical of the initial stage of the process, the nonlinear element may be replaced with an equivalent linear element. It is more difficult to solve the second problem because, in a more general sense, it involves an analysis of a feedback system under conditions when nonlinear effects may not be neglected.

Self-excitation of an elementary oscillator. Let us begin our study of the processes that take place in oscillators with what is known as a *transformer-coupled oscillator* (Fig. 14.12).

Referring to the figure, the oscillatory system is an LCR resonant circuit, and the feedback element is a coil L_{fb} so placed that the magnetic flux it sets up partly threads the coil L.

Suppose that in one way or another small oscillations have been excited in the circuit. If v is the voltage across the capacitor (and, in consequence, at the control electrode of the electron device), then, by the Kirchhoff second law, we may write the following differential

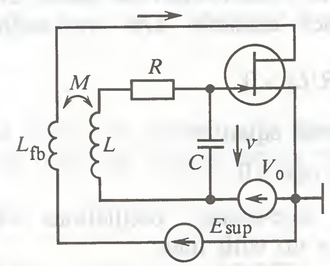


Fig. 14.12 Transformer-coupled oscillator

An oscillator can be built around a bipolar transistor on allowing for the additional attenuation caused in the oscillatory system by the finite output impedance

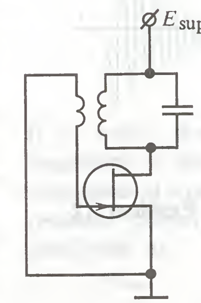
equation to describe the system in question:

$$LC\,d^2v/dt^2 + RC\,dv/dt + v = \pm M\,di/dt \tag{14.36}$$

where i is the current in the feedback circuit. The sign on the right-hand side of (14.36) is chosen according as the coils L and L_{fb} are connected in opposition or aiding.

Now we will make a basic assumption: Let the control voltage v be so small that the electron device may well be replaced with a controlled source of current which is a linear function of the control voltage:

$$i = i_0 + g_d v \tag{14.37}$$



Alternatively, the tuned circuit of a transformer-coupled oscillator may be connected in the output lead of the electron device

where i_0 is the direct component of current (of minor consequence for the subsequent discussion), and g_d is the dynamic (or incremental) transconductance defined as the slope of the current-voltage characteristic of the electron device at a selected operating point. On combining (14.36) and (14.37), we get

$$d^2v/dt^2 + (R/L \mp Mg_d/LC)\,dv/dt + \omega_0^2 v = 0 \tag{14.38}$$

where $\omega_0 = 1/\sqrt{LC}$ is the natural frequency of the lossless resonant circuit. It is important to note that Eq. (14.38) is a *linear* differential equation with constant coefficients. By varying the mutual inductance M, we can change the coefficient of the derivative dv/dt at will. The sign and magnitude of the coefficient determine, as will be recalled, the character of the free response of (or natural oscillations in) this dynamic system. If we choose the upper signs in Eqs. (14.36) and (14.38), feedback will lead to regeneration (already examined previously). If we let M reach its *critical* value, M_{cr},

$$M_{cr} = RC/g_d = 1/\omega_0 Q g_d \tag{14.39}$$

The critical value of mutual inductance

where Q is the Q-factor of the resonant circuit without regeneration, then Eq. (14.38) takes the form

$$d^2v/dt^2 + \omega_0^2 v = 0$$

which is typical of an ideal lossless oscillatory system. If $M > M_{cr}$, the system becomes unstable. On introducing the parameter

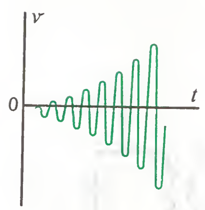
$$\alpha = \frac{1}{2}(Mg_d/LC - R/L) > 0$$

we get the differential equation

$$d^2v/dt^2 - 2\alpha dv/dt + \omega_0^2 v = 0$$

which describes sine-cosine oscillations whose amplitude exponentially builds up with time:

$$v(t) = A\exp(\alpha t)\cos\sqrt{\omega_0^2 - \alpha^2}\,t + B\exp(\alpha t)\sin\sqrt{\omega_0^2 - \alpha^2}\,t \quad (14.40)$$



▲ **Solve Problem 12**

where A and B are constants which depend on the initial conditions. Practically, $\alpha \ll \omega_0$ always, so, in accordance with Eq. (14.40), the carrier frequency of the oscillations produced in the linear mode is very close to the natural frequency of the resonant circuit.

It is important to stress the physical significance of the correct choice of the sign in Eq. (14.38): For the system to be capable of self-excitation, it is essential that any perturbation of its state should produce a feedback signal and that the feedback signal should combine with the initial perturbation so as to augment it. This is, in fact, the definition of positive feedback as it is treated in the theory of oscillatory systems.

Three-terminal oscillators. A disadvantage of a transformer-coupled oscillator is that it needs two inductors. In practice, use is more often made of circuits in which the feedback voltage is picked off an intermediate tap on the tank circuit. Quite aptly, they are called *three-terminal oscillators*. In the form shown in Fig. 14.13, usually called the *Hartley oscillator*, the intermediate tap is located on the inductive element of the tank.

Let us analyse the condition for the self-excitation of the Hartley oscillator. We will do this by developing and solving the characteristic equation for the circuit with its feedback loop closed.

If V_{in} and V_{out} are the Laplace transforms of the input and output signals of the system with its feedback loop open (Fig. 14.13*b*),

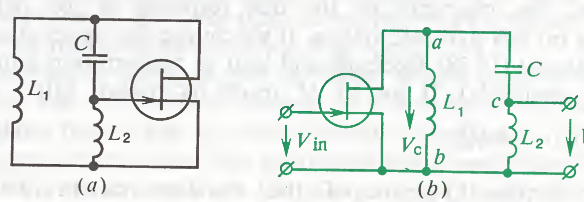


Fig. 14.13 Hartley oscillator: (*a*) schematic diagram (with the power and bias supplies omitted); (*b*) same, with the feedback loop open

and the transfer function of the system is $K(p)$, then the characteristic equation describing the closed-loop system has been found to be

$$K(p) = 1 \quad (14.41)$$

In order to find $K(p)$, we should note that, for the assumed direction of current flow, the voltage across the tuned circuit is

$$V_{ckt} = -g_d Z_{ab}(p) V_{in}$$

where

$$Z_{ab}(p) = k_{td}^2 Z_{ckt}(p)$$

Here, $k_{td} = L_1/(L_1 + L_2)$ is the tapping-down factor for the tank, and $Z_{ckt}(p)$ is the input impedance of the parallel tank between points a and c, that is, with connection across the whole of the tuned circuit. Deeming the tuned-circuit Q sufficiently high, we obtain the following approximate equation:

It is noted that the magnitude of the input impedance of a parallel resonant circuit is a maximum at $p = \pm j\omega_{res}$

$$Z_{ckt}(p) = R_{res}\left(\frac{1}{1 + p\tau_{ckt} - j\omega_{res}\tau_{ckt}} + \frac{1}{1 + p\tau_{ckt} + j\omega_{res}\tau_{ckt}}\right) \quad (14.42)$$

Finally, we note that at the resonant frequency

$$\omega_{res} = 1/\sqrt{(L_1 + L_2)C}$$

current resonance will exist in the tuned circuit and so

$$V_{out} = -(L_2/L_1)V_{ckt} \quad (14.43)$$

On combining the above equations, we may re-write the characteristic equation (14.41) as

$$(L_2/L_1)k_{td}^2 g_d Z_{ckt}(p) = 1 \quad (14.44)$$

or

$$\frac{1}{1 + p\tau_{ckt} - j\omega_{res}\tau_{ckt}} + \frac{1}{1 + p\tau_{ckt} + j\omega_{res}\tau_{ckt}} = \frac{1}{ag_d R_{res}} \quad (14.45)$$

where a is the coupling parameter defined as

$$a = L_2L_1/(L_1 + L_2)^2 \quad (14.46)$$

● **The condition for the self-excitation of a three-terminal oscillator**

Equation (14.45) has two complex conjugate roots:

$$p_{1,2} = \frac{ag_d R_{res}^{-1}}{\tau_{ckt}} \pm j\sqrt{\omega_{res}^2 - \frac{(ag_d R_{res})^2}{\tau_{ckt}^2}} \quad (14.47)$$

Thus, it is seen that the Hartley oscillator will jump into

self-excited oscillation subject to the condition

$$ag_d R_{\text{res}} > 1 \tag{14.48}$$

Example 14.5. *The tank circuit of a Hartley oscillator is tuned to $\omega_{\text{res}} = 6 \times 10^6\ \text{s}^{-1}$. The dynamic transconductance of the electron device at the operating point is $g_d = 7$ mA/V. The tank has $Q = 40$, the tank capacitance is $C = 400$ pF. Find the values of L_1 and L_2 necessary for the oscillator to jump into and sustain oscillations.*

The characteristic resistance of the tank circuit is

$$\rho = 1/\omega_{\text{res}} C = 416.6\ \Omega$$

Hence, the resonance resistance is

$$R_{\text{res}} = \rho Q = 16.6\ \text{k}\Omega$$

On the basis of Eq. (14.48), for the oscillator to jump into oscillation it is necessary that

$$a > 1/g_d R_{\text{res}} = 8.57 \times 10^{-3}$$

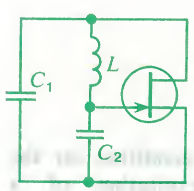
Allowing some margin, we take it that $a = 10^{-2}$. The total inductance of the tank circuit is

$$L_1 + L_2 = 1/\omega_{\text{res}}^2 C = 69.4\ \mu\text{H}$$

Using Eq. (14.46), we obtain a quadratic equation by which we can find L_2:

$$L_2(69.4 - L_2)/(69.4)^2 = 10^{-2}$$

which has two roots: $L_2 = 0.7\ \mu$H and $L_2 = 68.7\ \mu$H. From practical considerations, preference is given to the lower value of L_2 at which the oscillator will obviously develop a higher voltage across the tank circuit.



Another form of the three-terminal oscillator is the *Colpitts oscillator* in which the feedback voltage is picked off the common connection of two voltage-dividing capacitors C_1 and C_2. The condition for the self-excitation of the Colpitts oscillator is analysed along the lines similar to those for the Hartley oscillator.

RC phase-shift oscillators*. At frequencies below several tens of kilohertz it is increasingly more difficult to use *LC* resonant circuits as tanks for oscillators, mainly because the inductive elements become too large in size and weight. It is usual, therefore, to employ *RC* oscillators which are combinations of active two-ports (amplifiers) and passive *RC* feedback networks.

* Otherwise called *RC* harmonic oscillators.– *Translator's note.*

The condition imposed on the roots of the characteristic equation

Let $K(p)$ be the open-loop transfer function and $K(p) = 1$ be the characteristic equation describing the behaviour of a system with its feedback loop closed. For the system to be unstable and capable of producing harmonic oscillations in a steady state, it is necessary that the characteristic equation should have at least one pair of complex conjugate roots with a positive real part. The imaginary part of the roots will then define the frequency of oscillation.

Let us derive the condition for the self-excitation of a widely used oscillator type with two *RC* networks (Fig. 14.14). The basic

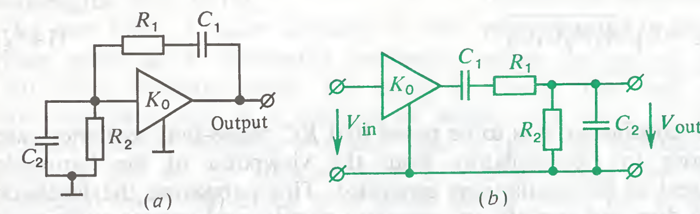


Fig. 14.14 Oscillator with two *RC*-networks: (*a*) schematic diagram; (*b*) same with the feedback loop open

element of the oscillator is an ideal amplifier which has a real and positive gain constant K_0. The amplifier output is connected to its input via a passive *RC* two-port whose transfer function in accordance with Fig. 14.14*b* has the form

$$K_1(p) = \frac{p\tau_1}{(1 + p\tau_1)(1 + p\tau_2) + p\tau'} \tag{14.49}$$

where $\tau_1 = R_1C_1$, $\tau_2 = R_2C_2$, and $\tau' = R_2C_1$. The characteristic equation of the oscillator, $K_0K_1(p) = 1$, may obviously be written as

$$p^2\tau_1\tau_2 + p[\tau_1 + \tau_2 - \tau'(K_0 - 1)] + 1 = 0 \tag{14.50}$$

This requirement stems from the Routh-Hurwitz criterion

The system becomes unstable when the coefficient of p crosses zero. In other words, for the oscillator to jump into self-sustained oscillations it is necessary that

$$\tau_1 + \tau_2 - \tau'(K_0 - 1) < 0$$

Hence follows the condition to be satisfied by the gain constant of the active element:

$$K_0 > 1 + \frac{R_1C_1 + R_2C_2}{R_2C_1} \tag{14.51}$$

Strictly speaking, the frequency of oscillation depends on the gain of the active element

Notably, if the two RC-networks are the same, the system will jump into oscillation at $K_0 > 3$.

The imaginary part of the roots of Eq. (14.50) depends not only on the circuit parameters R_1, R_2, C_1 and C_2, but also on the gain constant K_0. To determine the frequency of oscillation, it may be approximately taken that the oscillator is operating at the boundary of excitation, and so the coefficient of p is equal to zero. Then from the characteristic equation

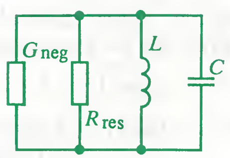
$$p^2\tau_1\tau_2 + 1 = 0$$

the frequency of oscillation is found to be

$$\omega_{osc} = 1/\sqrt{R_1R_2C_1C_2} \tag{14.52}$$

▲ Solve Problem 13

In conclusion, it is to be noted that RC phase-shift oscillators are inferior to LC-oscillators from the viewpoint of the harmonic content of the oscillations generated. This is because the feedback loop does not contain any resonant circuits and cannot sufficiently attenuate unwanted harmonics. An acceptable waveform of the resultant oscillations is secured through variations in the circuit design, such as the use of an additional nonlinear feedback loop with lag [37].



Internal-feedback oscillators. The types of oscillators examined above include suitably designed positive feedback loops. However, it is possible to build oscillators on a different principle, namely by introducing a negative conductance in the tank circuit.

If, for example, R_{res} is the tank resistance at resonance and $G_{neg} = g_d < 0$ is the parallel negative conductance corresponding to a small amplitude of oscillations, then the condition for the self-excitation of the system consists in making up for tank-circuit loss (see Chap. 8):

$$-g_d > 1/R_{res} \tag{14.53}$$

● **The power balance around an oscillator**

Let this condition be satisfied. As the oscillations build up owing to the nonlinear element, the rate of build-up slows down. In the steady state the power dissipated in the tank circuit over a period of natural oscillations is exactly equal to the power fed into the tank circuit from external sources over the same time interval. The self-regulation of the steady-state amplitude by this mechanism is known as *internal feedback*.

● **Internal feedback**

Distributed-parameter oscillators. Internal-feedback oscillators include a very interesting and practically very important variety built around sections of distributed-parameter transmission lines. A typical example is the *laser*, the name being derived from the initial letters of "Light Amplification by Stimulated Emission of

● **The laser**

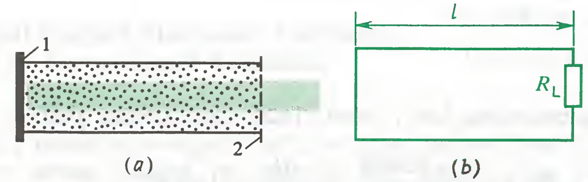


Fig. 14.15 Explaining the operation of a laser: (a) sketch; (b) equivalent circuit; (1) reflecting mirror; (2) semitransparent mirror

Radiation." This is a harmonic oscillator operating at optical wavelengths (Fig. 14.15).

■ **The operating principle of a laser**

In one form, a laser consists of two plane-parallel mirrors which make up a distributed oscillatory system known as the Fabry-Perot resonant cavity. The cavity is usually filled with an active medium, such as a mixture of helium and neon. In this medium, external pumping produces what is known as population inversion, that is, the preponderance of excited atoms over the nonexcited ones. On falling to a lower energy level, the excited atoms emit quanta of electromagnetic energy at a frequency equal to one of the resonant frequencies of the oscillatory system. If the emitted power exceeds the power delivered to load through a semitransparent mirror, the system jumps into self-sustained oscillation.

In order to get a quantitative idea about the condition for the self-excitation of a laser, consider its simple one-dimensional model shown in Fig. 14.15*b*. It consists of a short-circuited transmission-line section loaded into a resistor R_L. As the theory of transmission lines tells us [26], the harmonic wave process in a line is described by a complex propagation constant

Alternatively, this system may be described, assuming that the per unit length resistance is negative

$$\gamma = \alpha + j\beta = \sqrt{(R_1 + j\omega L_1)(G_1 + j\omega C_1)} \tag{14.54}$$

which involves the per unit length parameters R_1, G_1, L_1 and C_1. By the theory of distributed parameter circuits, the effect of pumping is customarily interpreted as rendering the per unit length shunt conductance negative: $G_1 = -g_1$. Then, deeming for simplicity that $R_1 = 0$ and that the conductance only slightly affects the phase velocity, that is, $g_1 \ll \omega C_1$, on the basis of (14.54) we have

$$\gamma \approx -\frac{1}{2}g_1 Z_w + j\beta_0 \tag{14.55}$$

where $Z_w = \sqrt{L_1/C_1}$ is the wave impedance of the line, and $\beta_0 = \omega\sqrt{L_1C_1}$ is the phase constant.

The load resistor is connected in parallel with the input admittance of the short-circuited line section of length l, equal, as

will be recalled, to

$Y_{in} = (1/Z_W)\coth\gamma l$

On substituting for γ from (14.55) and noting that at $\alpha \ll \beta$

$$\coth(\alpha l + j\beta l) \approx \frac{\alpha l - j\cot\beta l}{1 - j\alpha l\cot\beta l} \approx \frac{\alpha l}{\sin^2\beta l} - j\cot\beta l$$

we get

$$Y_{in} = \frac{1}{Z_W}\left(\frac{-g_1 l Z_W}{2\sin^2\beta_0 l} - j\cot\beta_0 l\right) \tag{14.56}$$

There exist an infinite number of frequencies at which a distributed-constant oscillatory system is capable of self-excitation

For the system to become unstable, it is required that the negative conductance of the line should balance the positive load conductance. This leads us to the condition for self-excitation

$$g_1 l/(2\sin^2\beta_0 l) > 1/R_L \tag{14.57}$$

The frequency of oscillation, ω_{osc}, is found from the resonance condition according to which the reactive component of the input impedance of the line must vanish:

$\cot\beta_0 l = 0$

Hence,

$\beta_0 l = (\pi/2)(2k+1) \quad k = 0, 1, 2, \ldots$

Therefore,

$$\omega_{osc} = \frac{\pi}{2l\sqrt{L_1 C_1}}(2k+1)$$

At the frequency of oscillation,

$|\sin\beta_0 l| = 1$

so the condition for oscillation takes the form

The condition for the self-excitation of a distributed-constant oscillatory system

$g_1 > 2/lR_L$

As follows from the above expression, the excitation of a distributed-parameter oscillator of the type in question becomes progressively easier with an increase in the load impedance, that is, as the line comes closer to the open-circuit condition with respect to the output terminals.

14.5 Self-Excited Harmonic Oscillators. The Large-Signal Condition

This section will outline the theory of oscillatory systems operating under large-signal conditions when the nonlinearity of the electron device cannot be ignored any longer. The material presented is based on an approximate solution of the nonlinear differential equation of the oscillator.

The abridged equation method. We go back to the simplest transformer-coupled oscillator and assume that the current-voltage characteristic, $i = f(v)$, of the active element is specified or known in advance. Since

$di/dt = (df/dv)(dv/dt)$

we may write the nonlinear differential equation (14.36) describing the behaviour of the oscillator under any conditions as

The choice of the sign for M assures the self-excitation of the system

$$\frac{d^2v}{dt^2} + \left(\frac{R}{L} - \frac{M}{LC}\frac{df}{dv}\right)\frac{dv}{dt} + \omega_0^2 v = 0 \tag{14.58}$$

Methods which would yield an exact solution for an equation like that are nonexistent. As a rule, one has to resort to simplifying assumptions of physical character and seek approximate solutions. In the case on hand, it is useful to note that the oscillator contains a high-Q tank, so, despite its nonlinearity, the voltage across the tank must resemble a harmonic wave at frequency ω_0 very closely.

Let us seek an approximate solution for Eq. (14.58) in the form

$v(t) = V(t)\cos\omega_0 t$

such that the amplitude $V(t)$ is assumed to be a slow function of time in the sense that

$|dV/dt| \ll \omega_0 |V|$

On this basis, in the expression for the derivative

$$\frac{dv}{dt} = \frac{dV}{dt}\cos\omega_0 t - \omega_0 V\sin\omega_0 t$$

we may retain only the second term

The derivative d^2V/dt^2 is small due to the assumption that the amplitude varies slowly

$$\frac{dv}{dt} \approx -\omega_0 V\sin\omega_0 t \tag{14.59}$$

Similarly,

$$\frac{d^2v}{dt^2} = \frac{d^2V}{dt^2}\cos\omega_0 t - 2\omega_0\frac{dV}{dt}\sin\omega_0 t$$

The abridged equation of an oscillator

$$-\omega_0^2 V\cos\omega_0 t \approx -2\omega_0\frac{dV}{dt}\sin\omega_0 t - \omega_0^2 V\cos\omega_0 t \qquad (14.60)$$

On substituting (14.59) and (14.60) into (14.58), we get what is known as the *abridged differential equation*

$$\frac{dV}{dt} + \frac{1}{2}\left(\frac{R}{L} - \frac{M}{LC}\frac{df}{dv}\right)V = 0 \qquad (14.61)$$

which approximately describes the events taking place in an oscillator using a high-Q tank circuit.

The abridged equation method simplifies the subsequent steps of analysis, because Eq. (14.61) contains only a first-order derivative.

The mean transconductance. The derivative df/dv in the second term of Eq. (14.61) is the local value of the dynamic (incremental) admittance of the nonlinear element. Since we assume that the output signal of the oscillator only slightly differs from a harmonic wave at frequency ω_0, then the current $i(t)$ is a periodic time function expandable into a Fourier series:

$$i = f(v) = I_0 + I_1\cos\omega_0 t + I_2\cos 2\omega_0 t + \dots$$

On discarding the harmonics, we get

$$f \approx I_0 + I_1\cos\omega_0 t$$

On the other hand,

$$v(t) = V\cos\omega_0 t$$

and so

$$\frac{df}{dv} = \frac{df}{dt}\bigg/\frac{dv}{dt} = I_1/V$$

The mean transconductance

By definition, the coefficient of proportionality between the fundamental amplitude of the current and the amplitude of the voltage at the control electrode is the *mean*, or *fundamental, transconductance* g_{m1}:

$$g_{m1}(V) = I_1(V)/V \qquad (14.62)$$

On inserting the mean transconductance, we may re-write the

abridged equation (14.61) as

$$\frac{dV}{dt} + \frac{1}{2}\left[\frac{R}{L} - \frac{M}{LC}g_{m1}(V)\right]V = 0 \qquad (14.63)$$

From an analytical point of view, it is more convenient when the current-voltage characteristic has the form of a power series:

Solve Problem 14

$$i(v) = a_0 + a_1 v + a_2 v^2 + \dots$$

The concept of mean transconductance was introduced by Yu. B. Kobzarev (USSR) who formulated a quasilinear theory of self-excited oscillators in the 1930s

Then, as will be recalled (see Chap. 11),

$$I_1 = a_1 V + {}^3/_4 a_3 V^3 + {}^5/_8 a_5 V^5 + \dots$$

so that

$$g_{m1}(V) = a_1 + {}^3/_4 a_3 V^2 + {}^5/_8 a_5 V^4 + \dots \qquad (14.64)$$

On substituting the above expression of the mean transconductance in the abridged equation (14.63), we obtain a 1st-order nonlinear differential equation which can always be solved by separation of variables.

Steady-state operation of the oscillator. By definition, the steady-state output signal of an oscillator has a constant amplitude. On setting $dV/dt = 0$ in Eq. (14.63), we get

The equation defining the steady-state amplitude of self-excited oscillations

$$g_{m1}(V) = RC/M \qquad (14.65)$$

the positive roots of which define the steady-state amplitude of oscillation.

Example 14.6. *Consider an oscillator in which the active element has the following mean transconductance as a function of the amplitude of control voltage:*

$$g_{m1}(V) = a_1 + {}^3/_4 a_3 V^2$$

where $a_1 = g_d = 15$ mA/V, *and* $a_3 = -4$ mA/V^3.

The parameters of the oscillator are as follows:

$$\omega_0 = 10^7\ \text{s}^{-1},\quad Q = 30,\ \text{and}\quad M = 1\ \mu\text{H}$$

Before we can determine the steady-state amplitude, we should first find

Solve Problem 15

$$\xi = RC/M = 1/\omega_0 QM = 3.33\times 10^{-3}\ \text{S}$$

On the basis of Eq. (14.65), the steady-state amplitude in our case satisfies the quadratic equation

$$a_1 + {}^3/_4 a_3 V_{ss}^2 = \xi$$

On solving it, we get

$$V=\sqrt{\frac{\xi-a_1}{{}^3/_4 a_3}}=1.97\text{ V}$$

Soft and hard self-excitation

The characteristic $g_{m1}(V)$ may have any one of the two forms shown in Fig. 14.16, depending on where the operating point of the nonlinear element is located.

If the mean transconductance monotonically decreases with increasing amplitude of the control voltage, the oscillator is said to be operating with *soft self-excitation.* The respective plot is shown in Fig. 14.16*a* which also shows the so-called feedback line–a horizontal line whose ordinate is RC/M. The intersection of the $g_{m1}(V)$ curve and the feedback line defines the only amplitude of steady-state oscillation, V_{ss}.

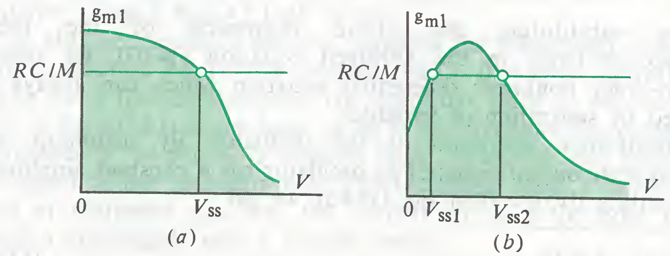


Fig. 14.16 Typical plots of mean transconductance versus the amplitude of control voltage: (*a*) "soft mode" characteristic; (*b*) "hard mode" characteristic

The situation is more complicated when the oscillator is operating with what is called *hard self-excitation.* Here, as is seen from the plot in Fig. 14.16*b*, two steady states are possible, one with amplitude V_{ss1}, and the other with amplitude V_{ss2}.

Stability of steady states. In order to see which of the two steady states is actually effected, we must define the crucial matter of their stability. An oscillatory process is said to be in a stable steady state if, after any small deviations of the amplitude of harmonic oscillations from the steady-state value, the system tends to go back to the state in which the amplitude is steady. Conversely, the process associated with an unstable steady state will tend to change its amplitude so as to move to another stable point.

The difference between the stability of linear and nonlinear oscillatory systems

The stability of a steady state is a concept specific of nonlinear oscillatory systems. It should be recalled that with reference to a linear dynamic system, we may only speak of stability or instability in the quiescent state (the small oscillation condition).

Consider the abridged equation (14.63) of the oscillator and assume that the amplitude V of self-sustained oscillations has departed by a small amount U from the steady-state value:

$$V=V_{ss}+U \tag{14.66}$$

Then

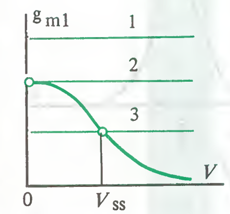
$$g_{m1}(V)\approx g_{m1}(V_{ss})+AU$$

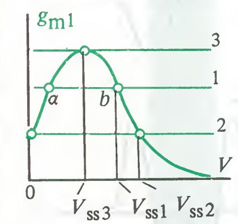
where $A=dg_{m1}/dV$ is the slope of the mean transconductance curve at the steady-state point. On substituting (14.66) into (14.63), we get, subject to Eq. (14.65), a differential equation in the amplitude increment

$$\frac{dU}{dt}=\frac{AM}{2LC}(U^2+UV_{ss})\approx\frac{AMV_{ss}}{2LC}U \tag{14.67}$$

The steady-state stability criterion

This simple linear differential equation points out that the sign of the derivative dU/dt depends on the sign of A. Thus, if $A<0$, then U and dU/dt take different signs. Therefore, if, for one reason or another, the amplitude V of the oscillation has become greater than V_{ss}, that is, $U>0$, then, by virtue of Eq. (14.67), the derivative $dU/dt<0$, and so with time the oscillatory system will regain its steady state.





It is easy to see that an oscillator operating in the soft self-excitation mode does possess the above property. Suppose that the mutual inductance M is so small that the feedback line *1* does not intersect the $g_{m1}(V)$ curve. The only steady state available to the system now is the quiescent state where the oscillations have zero amplitude. If M is increased, then at $M_{cr}=RC/g_d$ (line *2*) the oscillator will excite itself, no matter how small the amplitude of steady-state oscillation may be. Any further increase in M will lead to a gradual increase in the amplitude of the resultant oscillations (line *3*).

The events take a different course in an oscillator operating in the hard self-excitation mode. If initially the system is in the quiescent state and the feedback line takes up position *1*, then no self-excitation will occur, despite the fact that there are two steady-state points, namely the unstable point *a* and the stable point *b*. If, however, some external sources drive the system into harmonic oscillations at resonant frequency and with an amplitude corresponding to point *a*, then, since $A>0$ (the mean transconductance increases with increasing amplitude), the resulting oscillations will be unstable. Their amplitude will build up only until the system has moved to the stable point *b* characterized by a constant steady-state amplitude V_{ss1}.

An oscillator with hard self-excitation is capable of jumping into self-sustained oscillations as well. For this to happen, the feedback line must take up position *2* in which the steady state with an infinitesimal amplitude of oscillations is unstable. As follows from the foregoing, the oscillations thus excited will keep building up

until their amplitude reaches in the limit the steady-state value V_{ss2}. If we now reduce M, the amplitude of oscillations will gradually diminish until it is equal to V_{ss3}, and the feedback line takes up position *3* in which it is tangent to the $g_{m1}(V)$ curve. Any further decrease in M results in quenching the self-excited oscillations; their amplitude suddenly drops to zero.

▲ **Solve Problem 16**

Thus, in an oscillator with hard self-excitation, oscillations occur and disappear at various values of the feedback factor. This property is called *oscillatory hysteresis*.

● **Oscillatory hysteresis**

Dependence of soft and hard self-excitation of an oscillator on the location of the operating point. As already noted, the soft self-excitation mode differs from the hard mode in that at small amplitudes of oscillation the mean transconductance $g_{m1}(V)$ decreases with increasing V in the former case and increases in the latter case. Let the current-voltage characteristic of the nonlinear element be described by a power series. Then (see Eq. (14.66)), at low values of V the following relation exists:

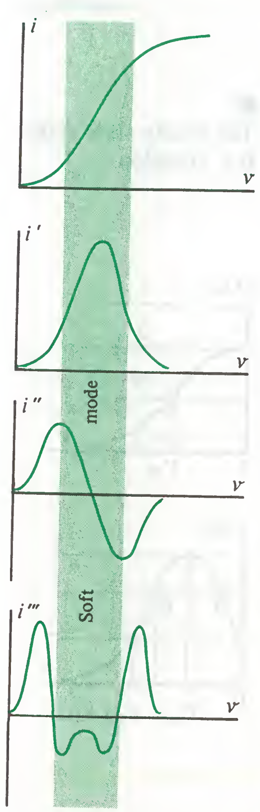
$$g_{m1}(V) = a_1 + {}^3/_4 a_3 V^2$$

from which it follows that the oscillator operates in the soft mode when $a_3 < 0$, or in the hard mode when $a_3 > 0$. As will be recalled,

$$a_3 = \frac{1}{3!}\frac{d^3 i}{dv^3}$$

with the derivative being evaluated at the point corresponding to the initial bias voltage applied to the nonlinear element.

As a rule, a plot of $i(v)$ yields a smooth curve monotonically increasing from zero to some constant level. If we differentiate this relation three times graphically, we will see that the soft mode is realized when the operating point is located in the middle portion of the characteristic.



The boundaries of soft and hard self-excitation

Transition to a steady-state amplitude. The abridged equation method makes it possible not only to find steady states and to analyse them for stability, but also to investigate the transition of the system to the steady-state amplitude of oscillations. To illustrate the method, let us find the manner in which the amplitude $V(t)$ varies with time in an oscillator operating in the soft mode, assuming that at $t = 0$ the system sustains harmonic oscillations at resonant frequency and some known amplitude V_0.

The problem reduces to solving the equation

$$\frac{dV}{dt} + \frac{1}{2}\left(\frac{R}{L} - \frac{Ma_1}{LC} - \frac{3a_3 M}{4LC} V^2\right) V = 0$$

subject to $V(0) = V_0$.

Let us introduce abbreviated notation:

$$\alpha = \frac{1}{2}\left(\frac{Ma_1}{LC} - \frac{R}{L}\right) > 0$$

$$\beta = \frac{3a_3 M}{8LC} < 0$$

We can multiply both sides of the abridged equation

$$\frac{dV}{dt} = (\alpha + \beta V^2) V \tag{14.68}$$

by V and obtain an equivalent form:

$$\frac{dV^2}{2(\alpha + \beta V^2) V^2} = dt \tag{14.69}$$

The events taking place in an oscillator with hard self-excitation are treated similarly

By expanding the left-hand side of (14.69) into partial fractions, we get

$$\frac{-(\beta/2\alpha)\, dV^2}{\alpha + \beta V^2} + \frac{dV^2/2\alpha}{V^2} = dt \tag{14.70}$$

Finally, on integrating subject to the initial condition, we obtain the solution of the problem as

$$\frac{1}{2\alpha}\ln\frac{V^2(\alpha + \beta V_0^2)}{V_0^2(\alpha + \beta V^2)} = t$$

Hence, the final result is

$$V(t) = \frac{V_0\sqrt{\alpha}\exp(\alpha t)}{\sqrt{\alpha + \beta V_0^2[1 - \exp(2\alpha t)]}} \tag{14.71}$$

As t tends to infinity, the amplitude of oscillations tends to a constant steady-state level

$$V_{ss} = \sqrt{\alpha / -\beta} = \sqrt{\frac{RC/M - a_1}{{}^3/_4 a_3}} \tag{14.72}$$

which checks with the result obtained in Example 14.6.

It is important to note that the steady-state amplitude of self-sustained oscillations does not depend on the initial conditions. If $V_0 = 0$, then, in accord with (14.71), the amplitude $V(t)$ is zero for any $t > 0$. However, since at $\alpha > 0$ the state of small oscillations is unstable, the self-excitation of the oscillator will always take place with any value of V_0, however small, arising, say, from thermal noise.

■ **The difference between active and passive oscillatory systems**

Lastly, it is to be noted that with increasing Q-factor of the

tank, when R tends to zero, the parameter α increases, too. This, in turn, implies that the transition to a steady-state amplitude proceeds faster at higher values of the tank Q (see Eq. (14.71)). It is useful to recollect that the passive narrowband systems studied in Chap. 9 have diametrically opposite properties. In their case, given a nonsteady process, the envelope varies progressively more slowly as the Q of the system is increased.

The phase-plane method. In radio engineering and radio physics, a special graphic method is used to describe and interpret the behaviour of oscillatory systems, notably oscillators. It is known as the *phase plane method* (or *phase plane analysis*). The rationale of the method can best be explained by taking the already studied 2nd-order linear oscillatory system as an example.

Free oscillations in the system are described by the equation

$$d^2x/dt^2 + 2\alpha\, dx/dt + \omega_0^2 x = 0 \qquad (14.73)$$

where the unknown function x may be a voltage, a current, etc. From the theory of differential equations it is known that by specifying x and dx/dt at any instant, we can fully describe the subsequent run of the process in the system. The graphical representation of the process will be the motion of a representative point along a curve (a *phase trajectory*) lying in the Cartesian x, x'-plane. This plane is called the *phase plane* of the system.

The phase portrait

For each set of initial conditions x and x' there is only one phase trajectory. If we take many different values of initial conditions, we will obtain an infinite number of nonintersecting phase trajectories which together form what is known as the *phase portrait* of the system.

The phase trajectories for Eq. (14.73) are fairly simple to construct because we know the general solution

$$x(t) = A\exp(-\alpha t)\cos(\omega_n t + \varphi)$$

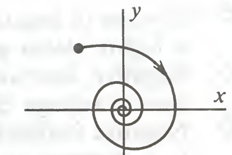
where A and φ are the constants dependent on the choice of the initial conditions; $\omega_n = \sqrt{\omega_0^2 - \alpha^2}$ is the natural frequency. For $\alpha \ll \omega_0$, we approximately have

$$dx/dt \approx -\omega_n A\exp(-\alpha t)\sin(\omega_n t + \varphi)$$

It is convenient to introduce the variable $y = (dx/dt)/\omega_n$. Then the oscillatory process in the system will be represented by two time-dependent projections of the representative point on the x- and the y-axes:

$$x(t) = A\exp(-\alpha t)\cos(\omega_n t + \varphi)$$
$$y(t) = -A\exp(-\alpha t)\sin(\omega_n t + \varphi) \qquad (14.74)$$

It is easy to see that Eqs. (14.74) describe a logarithmic spiral



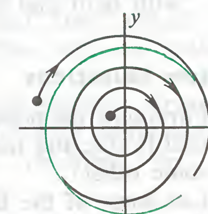
whose radius vector exponentially decreases in time as $A\exp(-\alpha t)$, if $\alpha > 0$ (the process is decaying). The representative point moves clockwise; the polar angle of the vector, $-\omega_n t - \varphi$, varies linearly with time. With t tending to infinity, the representative point spirals towards the origin, which corresponds to a state of stable equilibrium.

If $\alpha < 0$ (the process is building up), the phase trajectory spirals away from the origin, and this is an indication that the system is in a state of unstable equilibrium.

Finally, if $\alpha = 0$ (there is neither a build-up nor a decay of the process), then in our idealized system the representative point will be continuously moving round a circle whose radius depends on the choice of the initial conditions.

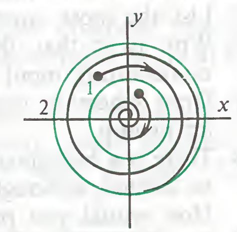
The limit cycle

Phase portraits of oscillators. Plotting phase trajectories for a nonlinear equation of the form in (14.58) which describes an oscillator may prove an extremely tedious problem solvable only by numerical or graphical methods. If, however, the Q-factor of the oscillator tank is high and the oscillations generated are close in waveform to harmonic, we will see that the phase plane contains a closed curve called a *limit cycle*, around which the representative point moves in the steady state. The radius of the limit cycle is equal to the amplitude of steady-state self-sustained oscillations and does not depend on the initial conditions.



The soft mode

Limit cycles may be stable and unstable. An oscillator operating in the soft self-excitation mode has only one limit cycle which is stable. To demonstrate, if initially the amplitude of oscillations is greater than its steady-state value, then, as time passes, the phase trajectory will spiral towards the limit cycle. By the same token, the amplitude of oscillations will build up if initially it was smaller than the value corresponding to the radius of the limit cycle.



The hard mode

The situation is more complicated for an oscillator with hard self-excitation. Now there are two limit cycles, *1* and *2*, of which cycle *1* has a smaller amplitude and is unstable while cycle *2* is stable. If initially the representative point is inside cycle *1*, the phase trajectory will spiral towards the origin. Therefore, some external force is needed in order to excite the oscillator. If, on the other hand, the representative point is located outside the unstable limit cycle *1*, the phase trajectory will spiral towards the stable limit cycle *2*.

Summary

- Feedback is widely used to build dynamic systems which will have a specified frequency or impulse response.
- There may be positive-feedback and negative-feedback systems.

- The use of negative feedback in an amplifier makes it possible to build systems with a highly stable gain.
- Negative feedback minimizes the nonlinear distortion arising from the nonlinear current-voltage characteristics of active elements.
- Negative feedback makes it possible to expand the bandwidth of an amplifier owing to a decrease in the gain level.
- The stability of small oscillations in feedback systems is analysed with the aid of stability criteria. Most frequently, these are the Routh-Hurwitz and Nyquist criteria.
- Active RC filters are frequency-selective devices containing no inductive elements.
- When positive feedback is applied to an amplifier, instability may arise. If the circuit contains an oscillatory system (a tank circuit), an unstable positive-feedback amplifier becomes a harmonic oscillator.
- Where the objective is to generate low-frequency harmonic oscillations, resort is made to RC phase-shift oscillators which do not contain any resonant (tank) circuits.
- In an oscillator the steady-state amplitude of oscillations is decided by the form of the nonlinearity displayed by its electron device.
- Oscillators may reside in a stable and an unstable steady state. Also, they may operate with hard and soft self-excitation.

Review Questions

1. Formulate the principle by which feedback is classified in dynamic systems. Is it possible that in one and the same system feedback is positive at one frequency, and negative at some other?
2. List some of the engineering applications of negative feedback in amplifiers.
3. What is the salient feature of the frequency response of an amplifier using delayed feedback?
4. Formulate the Routh-Hurwitz and Nyquist stability criteria.
5. List the most important properties of operational amplifiers.
6. Why is it that the connection of the output in an operational amplifier to its noninverting input disturbs its stability?
7. Draw schematic diagrams for the scale changer and the analog integrator built around an opamp.
8. There is a breadboard version of a transformer-coupled oscillator. The oscillator fails to be excited, although it has been assembled from the components known to be good. How would you proceed with the alignment of the oscillator?
9. Draw schematic diagrams for the Hartley and Colpitts oscillators.
10. Does the frequency of the oscillations generated by an RC oscillator depend on the gain constant of the active element?
11. What is the condition of oscillation for an internal-feedback oscillator?
12. Define the mean transconductance of an electron device.
13. What is the fundamental difference between the soft and hard self-excitation of an oscillator?
14. If the current-voltage characteristic of a nonlinear active element is known, how can you determine the boundary between the soft and hard modes of self-excitation?

Problems

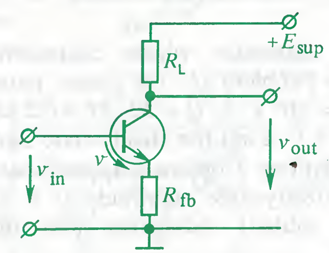
1. In a feedback system, the forward-path element has a frequency-independent gain $K_0 = 1000$. The gain of the feedback loop (feedback ratio) is

$$\beta_0 = 5 \times 10^{-4} \exp(-j45°)$$

Determine what kind of feedback, positive or negative, exists in the system.

2. An amplifier has a frequency-independent gain of $K_0 = 10\,000$. When the ambient conditions (say, the temperature) change, the gain changes too, such that $\Delta K_0/K_0 = 0.2$. The system parameters are stabilized by negative feedback. Find the value of the feedback ratio β for which the instability in the gain is reduced by a factor of 10, that is, to $\Delta K_{fb}/K_{fb} = 0.02$. What is the closed-loop (overall) gain of the system in the circumstances?

3. Demonstrate that in the amplifier shown in the accompanying diagram



there is negative feedback such that

$$\beta^0 = -R_{fb}/R_L$$

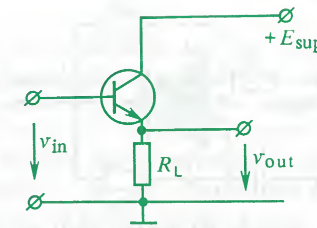
Hint: Use an auxiliary voltage v between the base and the emitter (see the diagram) and note that the a.c. component of emitter current is

$$i_E = g_m v$$

4. Find the voltage-ratio transfer function

$$K = v_{out}/v_{in}$$

of the emitter follower shown in the accompanying diagram:



Hint: The transconductance g_m of the transistor should be assumed known.

5. In a single-stage RC-loaded amplifier, $R_L = 1.6$ kΩ. The stray capacitance C_s in the collector lead is 50 pF. The transconductance of the transistor is $g_m = 30$ mA/V. The emitter lead contains a feedback resistor $R_{fb} = 27$ Ω (see Problem 3). Find the upper and lower cut-off frequencies of the amplifier with and without negative feedback.

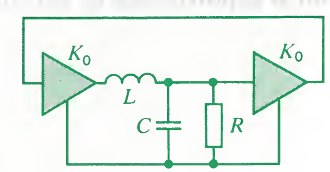
6. Using the Routh-Hurwitz criterion, analyse the stability of small oscillations in a dynamic system described by the characteristic equation:

(a) $p^4 + 3p^3 + 2p^2 + p + 1 = 0$

(b) $p^4 + 2p^3 + 3p^2 + p + 1 = 0$

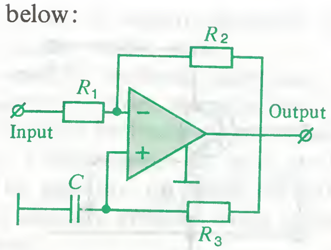
7. Using the Nyquist criterion, analyse a delayed-feedback system (see Fig. 14.4) for stability.

8. Using the Nyquist criterion, analyse for stability the closed-loop system containing two ideal amplifiers each of gain K_0:



9. Find the transfer function of the system

shown below:



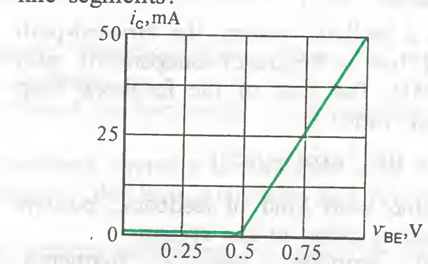
10. An analog integrator built around an opamp (see Fig. 14.9) has the following parameters: $R = 2.7$ kΩ, $C = 1.8$ nF. The opamp draws its power from a power supply with $E_{sup} = 12$ V. The signal applied to the integrator input is $v_{in} = 0.1\sigma(t)$. Find the output signal and determine the time required for the output voltage to reach the supply voltage level.

11. Synthesize a second-order Chebyshev active RC low-pass filter, the initial specifications being $\omega_c = 1.5 \times 10^2$ s^{-1} and $\varepsilon = 0.75$.

12. The transformer-coupled oscillator of Fig. 14.12 uses a tank circuit tuned to $f_{res} = 400$ kHz. The element values of the tank are $L = 15$ μH and $R = 8$ Ω. The dynamic transconductance of the transistor is $g_d = 1.5$ mA/V. Find the critical value of mutual inductance M assuring the self-excitation of the system.

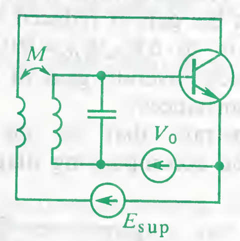
13. The RC oscillator of Fig. 14.14 generates harmonic oscillations at 250 Hz. Select the element values necessary for the oscillator to operate at the frequency stated, assuming that the oscillator is operating at the boundary of excitation.

14. The current-voltage characteristic of a transistor is approximated by two straight-line segments:



Construct a plot of the mean transconductance g_{m1} as a function of the amplitude of the high-frequency voltage V, assuming that the operating point is at $V_0 = 0.75$ V.

15. The harmonic oscillator shown schematically in the accompanying diagram



uses a transistor whose characteristic is given in Problem 14. The circuit parameters are $\omega_0 = 10^6$ s^{-1}, $Q = 50$, $M = 0.2$ μH, and $V_0 = 0.75$ V. Find the steady-state amplitude of oscillations. Construct approximate plots of the steady-state amplitude as a function of the mutual inductance and the bias voltage.

16. In the oscillator of the previous problem the initial bias is set at 0.25 V. Show that the system will operate with hard self-excitation. Find M at which the steady-state amplitude of oscillations is 1 V.

Advanced Problems

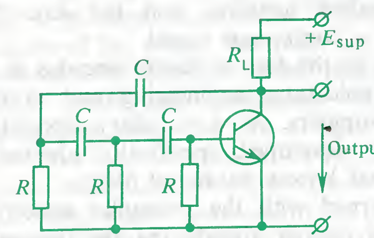
17. Show that the feedback system in Fig. 14.1 can be described by an infinite relation of the form

$$V_{out} = KV_{in} + \beta(KV_{in} + \beta(KV_{in} + \beta(KV_{in} + \ldots$$

relating the Laplace transforms of the input and output signals. Inquire into the equivalence between the above representation and Eq. (14.2). Discuss the treatment of the processes in a feedback system as the circulation of signals round the feedback loop.

18. Plot the Nyquist locus and, using the Nyquist stability criterion, investigate the stability of a three-stage RC-coupled amplifier, assuming that the output of the system is directly connected to the input.

19. Investigate the condition of self-excitation for a RC phase-shift oscillator set up as shown in the accompanying diagram:



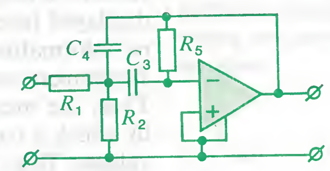
Derive an expression for the frequency of oscillation, neglecting the effect of the gain of the active element on the imaginary part of the roots of the characteristic equation.

20. Analyse the steady-state amplitude of oscillations in a distributed-parameter oscillatory system, assuming that the per-unit-length shunt conductance which describes the effect of the active medium is defined by

$$G_1 = -g_1 + \alpha V^2$$

where g_1 and α are constants, and V is the amplitude of h.f. oscillations.

21. Show that the transfer function of a bandpass filter is realized by the circuit of the form shown below:



Chapter 15

Discrete Signals. Principles of Digital Filtering

The difference between discrete and continuous (analog) signals has been underlined in Chapter 1 when considering the classification of signals. It is worth while to recollect the basic property of a discrete signal: Its values are defined at a countable set of points $(\ldots, t_0, t_1, t_2, \ldots)$ rather than for all instants of time. Therefore whereas an analog signal $x(t)$ is represented by a mathematical model possessing the usual properties of a smooth function, a discrete signal $x_d(t)$ is described by a series $(\ldots, x_0, x_1, x_2, \ldots)$ of its values at points $(\ldots, t_0, t_1, t_2, \ldots)$, respectively.

Discrete signals arise naturally in cases where a message source generates information at fixed instants of time. Typical discrete signals are the air temperature data transmitted by broadcasting stations several times a day. The salient feature of discrete signals is displayed here with utmost clarity: Between the broadcasts there is no information available about the quantity involved, although the air temperature is changing all the time in a quite gradual manner. Thus, the measured data are the outcomes of a special procedure by which a continuous quantity is converted into a series of discrete values. This procedure is called *sampling*, and the data thus generated are the *samples* of a continuous signal.

Discrete signals have come to the fore in recent decades in the wake of advances in communications engineering and in data processing on high-speed computers. As a corollary, specialized computing devices have been developed, specifically adapted to process digital information and known as *digital filters*.

This chapter will be concerned with the principles underlying a mathematical description of discrete signals and the theoretical foundations on which the synthesis of devices for digital data processing is based.

15.1 Discrete Pulse Sequences

For the first time discrete signals came into use in the 1940s with the advent of pulse-modulated communication systems. A salient feature of pulse modulation is the fact that the carrier is a periodic train of short pulses rather than a continuous harmonic wave.

Principle of pulse modulation. Consider the simple pulse modulator shown in block-diagram form in Fig. 15.1.

Referring to the figure, the pulse modulator is a device with two inputs. One input accepts the original analog signal $x(t)$ to be sampled. The other input is fed a periodic train of short

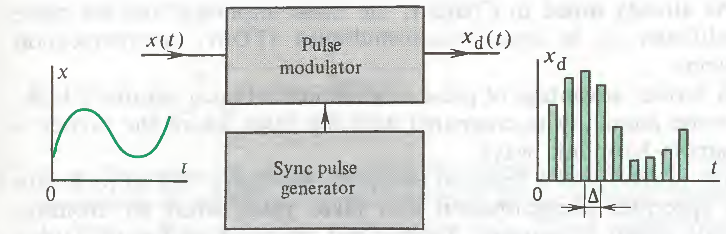


Fig. 15.1 Block diagram of a pulse modulator

synchronizing (sync) pulses equidistantly spaced in time, the spacing Δ being known as the *sampling interval*. The modulator is arranged so that each time a sync pulse is applied, the instantaneous value of $x(t)$ is measured, or sampled, and a short pulse whose area is proportional to the sample, appears at the output of the device.

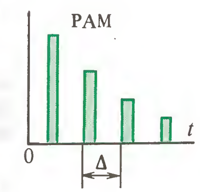
The sampling interval

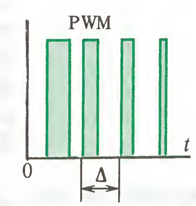
From the above principle, we may write the following mathematical model for the discrete, or sampled, waveform produced by pulse modulation:

With this model, the values of the signal between sampling instants are assumed to be zero

$$x_d(t) = \sum_{k=-\infty}^{\infty} x(k\Delta)\,\delta(t - k\Delta) \tag{15.1}$$

It is to be noted that from a formal mathematical point of view the shape of the pulses that make up the output signal is entirely immaterial. For example, these pulses may all be of the same duration while their amplitude varies in proportion to the instantaneous values of the original waveform at the sampling points. This form of modulation has come to be known as *pulse-amplitude modulation* (PAM for short). There is another procedure called *pulse-width*, *pulse-duration* or *pulse-length* modulation, (abbreviated PWM, PDM or PLM), in which the output pulses are of a constant amplitude, but their width (duration or length) is directly proportional to the instantaneous values of the analog signal.





The choice of a particular form of pulse modulation is dictated by engineering considerations (ease of physical implementation) and by the character of the signals being transmitted. For example, the use of PAM may be unwarranted if the original modulating signal varies between broad limits or, as is usually said, it has a broad dynamic range. The point is that a faithful transmission of such a signal would require a transmitter in which a rigorously linear relationship is maintained between the amplitudes of the input and output signals. This constraint is nonexistent in PWM systems, but they are more difficult to implement engineeringly than PAM system.

Use of PM systems

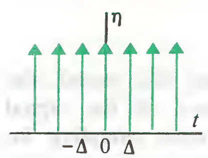
As already noted in Chap. 1, the most important use for pulse modulation is in time-division-multiplex (TDM) communication systems.

A further advantage of pulse modulation is that it assures a higher noise immunity as compared with the cases where the carrier is a simple harmonic wave.

The sampling sequence

The spectrum of a digitized (sampled) waveform. Let us look into the spectrum transformation that takes place when an arbitrary analog signal is sampled. To this end, we refer to Eq. (15.1) and note that the digitized (sampled) signal $x_d(t)$ is the product of the original wave $x(t)$ and the so-called *sampling sequence*

$$\eta(t) = \sum_{k=-\infty}^{\infty} \delta(t - k\Delta) \tag{15.2}$$



formed by delta impulses following at equal time intervals Δ. As will be recalled, the spectrum of the product of two signals is the convolution of their individual spectra (see Chap. 2). Therefore, since we know how the signals and their spectra are related to one another: $x(t) \leftrightarrow S_x(\omega)$ and $\eta(t) \leftrightarrow S_\eta(\omega)$, we may conclude on the basis of Eq. (2.35) that the spectrum of the sampled waveform, $S_{xd}(\omega)$, is equal to

$$S_{xd}(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} S_\eta(\omega) S_x(\omega - \xi)\, d\xi \tag{15.3}$$

In order to find the spectrum of the sampling sequence, $S_\eta(\omega)$, we expand the periodic function $\eta(t)$ into a Fourier series:

$$\eta(t) = \sum_{n=-\infty}^{\infty} C_n \exp(j\,2\pi n t/\Delta)$$

The coefficients of the series (the Fourier coefficients) are given by

$$C_n = \frac{1}{\Delta} \int_{-\Delta/2}^{\Delta/2} \delta(t) \exp(-j\,2\pi n t/\Delta)\, dt = 1/\Delta$$

The spectrum of the sampling sequence

Referring to Eq. (2.42), we get

$$S_\eta(\omega) = \frac{2\pi}{\Delta} \sum_{n=-\infty}^{\infty} \delta(\omega - 2\pi n/\Delta) \tag{15.4}$$

That is, the spectrum of the sampling sequence consists of an infinite series of delta impulses displaced along the frequency axis so that they are separated by an interval equal to $2\pi/\Delta$ (s^{-1}).

Finally, on substituting (15.4) into (15.3) and reversing the order

of summation and integration, we get

$$S_{xd}(\omega) = \frac{1}{\Delta} \sum_{n=-\infty}^{\infty} S_x(\omega - 2\pi n/\Delta) \tag{15.5}$$

Thus, the spectrum of the sampled waveform consists of an infinite succession of repetitions of the spectrum of the original waveform (to within the negligible scale factor $1/\Delta$). These repetitions are displaced along the frequency axis so that they are separated by the interval $2\pi/\Delta$, equal to the angular sampling rate (Fig. 15.2).

Solve Problem 1

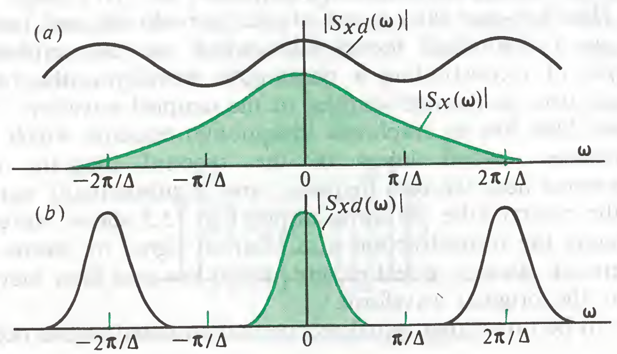


Fig. 15.2 Spectra (green areas) of the sampled waveform for two values of the upper frequency limit in the signal spectrum: (*a*) upper frequency limit exceeds half the sampling rate; (*b*) upper frequency limit is less than half the sampling rate

The reconstruction of a continuous waveform from a sampled waveform. If ω_u is the upper frequency limit in the spectrum of the original waveform, then, as can be seen from Fig. 15.2, for $\omega_u \leqslant \pi/\Delta$, the individual lobes of the spectral diagram do not overlap any longer. Therefore, it is possible to reconstruct the original continuous waveform by passing the sampled waveform defined in (15.1) through a low-pass filter. The maximum allowable sampling interval will then be $\Delta = \pi/\omega_u = 1/2f_u$, which completely agrees with the sampling theorem.

Let the filter used to reconstitute the original continuous waveform have the frequency response function

$$K(j\omega) = \begin{cases} 0, & \omega < -\omega_u \\ K_0, & -\omega_u < \omega < \omega_u \\ 0, & \omega > \omega_u \end{cases}$$

Its impulse response is then

The amplitude factor has been chosen to make the result more instructive

$$h(t)=\frac{1}{2\pi}\int_{-\infty}^{\infty} K(j\omega)\exp(j\omega t)\,d\omega=(K_0\omega_u/\pi)\sin(\omega_u t)/\omega_u t$$

On setting $K_0=\pi/\omega_u$ and noting that the sampled waveform defined in (15.1) is the weighted sum of delta impulses, the waveform at the output of the reconstituting filter is found to be

$$y(t)=\sum_{k=-\infty}^{\infty} x(k\Delta)\frac{\sin\omega_u(t-k\pi/\omega_u)}{\omega_u(t-k\pi/\omega_u)} \tag{15.6}$$

The reconstruction filter may be a high-order Butterworth low-pass filter

which, in accord with (5.18), is an exact repetition of the original bandwidth-limited waveform $x(t)$.

An ideal low-pass filter is not physically realizable and can only serve as a theoretical model with which we can explain the principle of reconstituting a continuous waveform (the original message) from its discrete samples, or the sampled waveform. A real low-pass filter has an amplitude (magnitude) response which either encompasses several lobes of the spectral diagram or is concentrated near the zero frequency and is substantially narrower than the central lobe. As an example, Fig. 15.3 shows waveforms illustrating the reconstruction of a sampled signal by means of an RC network. As seen, a real reconstruction low-pass filter inevitably distorts the original waveform.

It is to be noted that signal reconstruction could utilize not only the central, but any side lobe of the spectral diagram. In addition to low-pass filtering, however, this would necessitate resort to frequency transformation or synchronous detection.

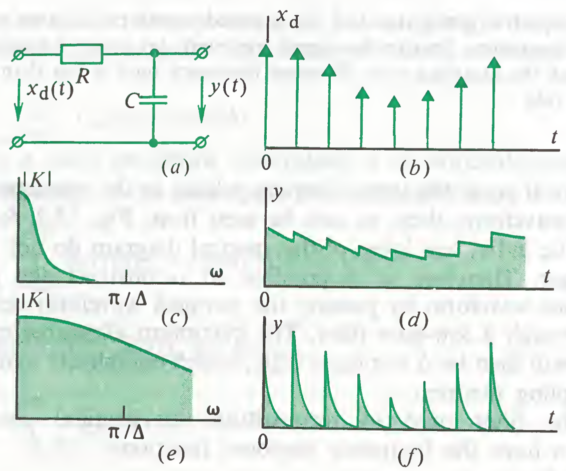


Fig. 15.3 Reconstruction of a continuous waveform from its samples with an RC-network: (a) filter schematic; (b) sampled waveform; (c and d) amplitude response of the filter and the signal at its output for $RC \gg \Delta$; (e and f) same for $RC \ll \Delta$

15.2 Digitization of Periodic Signals

In signal analysis by computers, the following situation is typical: A continuous signal $x(t)$ is specified over a time interval $(0, T)$ by its sampled values $(x_0, x_1, \ldots, x_{N-1})$ taken respectively at times $(0, \Delta, \ldots, (N-1)\Delta)$; the total number of sampled values is $N=T/\Delta$. The collection of these numbers, real or complex, is the only information from which we can judge about the spectral properties of the signal $x(t)$. The procedure by which such discrete signals are analysed consists in that the collection of sampled values is repeated mentally an infinite number of times. As a result, the signal becomes periodic (Fig. 15.4).

The samples shown in colour belong to the periodicity interval

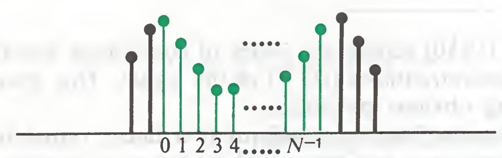


Fig. 15.4 Sampled representation of a periodic signal

On assigning to this signal a particular mathematical model, we can use an expansion into a Fourier series and find the corresponding amplitude (Fourier) coefficients. The collection of these coefficients forms the spectrum of the periodic discrete signal.

The discrete Fourier transform. Let us use a model in the form of a series of delta impulses and assign to the original waveform $x(t)$ its discrete representation on the inerval $(0, T)$:

$$x_d(t)=\sum_{k=0}^{N-1} x_k\delta(t-k\Delta) \tag{15.7}$$

where $x_k=x(k\Delta)$ are the sampled values at the kth points.

Now we represent the discrete model in (15.7) by a complex Fourier series

$$x_d(t)=\sum_{n=-\infty}^{\infty} C_n\exp(j2\pi nt/T) \tag{15.8}$$

in which the Fourier coefficients are defined as

$$C_n=\frac{1}{T}\int_0^T x_d(t)\exp(-j2\pi nt/T)\,dt \tag{15.9}$$

On substituting (15.7) into (15.9) and introducing a dimensionless variable $\xi = t/\Delta$, we obtain

$$C_n = \frac{1}{N\Delta}\int_0^{N\Delta}\sum_{k=0}^{N-1} x_k\delta(t-k\Delta)\exp(-j2\pi nt/T)\,dt$$

$$= \frac{1}{N}\int_0^{N}\sum_{k=0}^{N-1} x_k\delta(\xi-k)\exp(-j2\pi n\xi/N)\,d\xi$$

$$= \frac{1}{N}\sum_{k=0}^{N-1} x_k\int_0^{N}\delta(\xi-k)\exp(-j2\pi n\xi/N)\,d\xi$$

Solve Problem 2

Finally, by utilizing the filtering properties of the delta-function, we have

$$C_n = \frac{1}{N}\sum_{k=0}^{N-1} x_k\exp(-j2\pi nk/N) \qquad (15.10)$$

Properties of the discrete Fourier transform (DFT)

Equation (15.10) defines the series of coefficients which form the *discrete Fourier transform (DFT)* of the signal. This transform has the following obvious properties.

1. The discrete Fourier transform is a linear transform, that is, a sum of signals can be represented by the sum of their individual discrete Fourier transforms.

2. The number of Fourier coefficients $C_0, C_1, \ldots, C_{N-1}$ found by Eq. (15.10) is equal to the number N of sampled values in the collection. Thus, for $n = N$, $C_N = C_0$.

3. The coefficient C_0 (the constant component) is the mean value of all sampled values:

$$C_0 = \frac{1}{N}\sum_{k=0}^{N-1} x_k$$

4. If the number N of samples is even, then

$$C_{N/2} = \frac{1}{N}\sum_{k=0}^{N-1} x_k(-1)^k$$

5. If the sampled values x_k are real numbers, then the Fourier coefficients located symmetrically about $N/2$ form complex conjugate pairs:

$$C_{N-n} = \frac{1}{N}\sum_{k=0}^{N-1} x_k\exp[-j2\pi(N-n)k/N]$$

$$= \frac{1}{N}\sum_{k=0}^{N-1} x_k\exp(j2\pi nk/N) = C_n^*$$

Therefore it is legitimate to think that the coefficients $C_{N/2+1}, \ldots, C_{N-1}$ correspond to negative frequencies. In the analysis of the amplitude spectrum of a signal they do not play any role, so they need not be calculated.

Example 15.1. *Over its periodicity interval a sampled signal is specified by six equidistant samples $\{x_k\} = (1, 1, 1, 0, 0, 0)$. Find the coefficients of the discrete Fourier transform for this signal.*

Using Eq. (15.10), we directly calculate:

$$C_0 = \frac{3}{6} = 0.5$$

$$C_1 = \frac{1}{6}(1 + 1\times e^{-j\pi/3} + 1\times e^{-j2\pi/3}) = \frac{1}{6}(1 - j\sqrt{3})$$

$$C_2 = \frac{1}{6}(1 + 1\times e^{-j2\pi/3} + 1\times e^{-j4\pi/3}) = 0$$

$$C_3 = \frac{1}{6}(1 + e^{j\pi} + e^{j2\pi}) = \frac{1}{6}$$

The remaining coefficients are found on the basis of their conjugacy:

$$C_4 = C_2^* = 0, \quad C_5 = C_1^* = \frac{1}{6}(1 + j\sqrt{3})$$

The number of harmonics being found

Thus, since we have a sampled signal with $N = 6$ samples, we can find the constant component and the complex amplitudes of the fundamental, the second and third harmonics of the original continuous waveform. Clearly, with any even number N, the number of harmonics thus found account for half the sampled values. This stems directly from the Kotelnikov sampling theorem. To demonstrate, the upper frequency limit in the spectrum of the sampled waveform is $f_u = 1/(2\Delta) = (N/2)f_1$, where $f_1 = 1/T$ is the fundamental frequency.

The reconstruction of the original waveform from its discrete Fourier transform. Once the discrete Fourier coefficients (C_0, C_1, $C_2, \ldots, C_{N/2}$) have been found from the collection of samples (x_0, $x_1, \ldots, x_{N-1}$) of a real signal, we can always reconstruct from them the original waveform $x(t)$. Obviously, the Fourier series of such a signal takes the form of a finite sum

It is assumed in advance that the original signal satisfies the conditions of the Kotelnikov (sampling) theorem

$$x(t) = C_0 + 2|C_1|\cos(2\pi t/T + \varphi_1) + 2|C_2|\cos(4\pi t/T + \varphi_2)$$
$$+ \ldots |C_{N/2}|\cos(N\pi t/T + \varphi_{N/2}) \qquad (15.11)$$

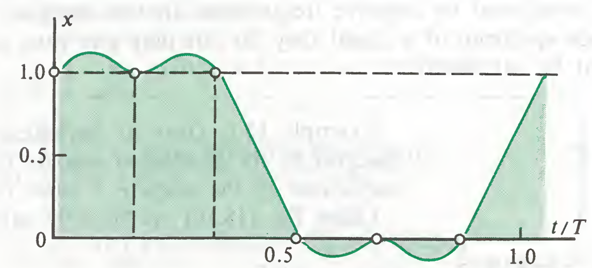


Fig. 15.5 Signal reconstituted from discrete Fourier transform coefficients

where $\varphi_i = \arg C_i$ are the phase angles of the corresponding Fourier coefficients. As an example, Fig. 15.5 shows a signal $x(t)$ reconstructed from its samples in accordance with the data given in Example 15.1. On the basis of Eq. (15.11), this signal may be defined as

$$x(t) = \frac{1}{2} + \frac{2}{3}\cos(2\pi t/T - \pi/3) + \frac{1}{6}\cos(6\pi t/T)$$

▲ Solve Problem 3

It should be especially stressed that the reconstruction of a continuous waveform by Eq. (15.11) *is not an approximate, but an exact operation* fully equivalent to obtaining the instantaneous values of a band-limited signal from its samples making up the Kotelnikov series. In some cases, however, it is preferable to use the procedure based on the discrete Fourier transform, because it leads to finite sums of harmonics, while the Kotelnikov series for a periodic signal must in principle contain an infinite number of members.

The inverse discrete Fourier transform (IDFT). The task of discrete spectral analysis may be stated in a different way. Suppose that the discrete Fourier coefficients C_n are specified in advance. Also let us set $t = k\Delta$ in Eq. (15.8) and note that the sum is taken only over a finite number of members in the series, which correspond to the harmonics contained in the spectrum of the original waveform. Hence, we obtain an equation for sampled values

$$x_k = \sum_{n=0}^{N-1} C_n \exp(j2\pi nk/N) \qquad (15.12)$$

which defines the algorithm of the *inverse discrete Fourier transform (IDFT)*.

The mutually complementing equations (15.10) and (15.12) are the discrete counterparts of the usual Fourier transform pair for continuous signals.

▲ Solve Problem 4

At present, discrete spectral analysis is one of the most widely used techniques of numerical analysis of signals on a computer. As an example, Appendix 5 gives a FORTRAN subroutine for computing the discrete Fourier transform. It should be noted that with a large number of sampled values, any direct calculation of the sum in (15.10) or (15.12) involves the expenditure of a lot of computer time. To avoid this, resort is widely made to what is called as the *fast Fourier transform (FFT)*. By purely algorithmic means, it substantially reduces the amount of computations involved in finding coefficients for the discrete Fourier transform and the inverse discrete Fourier transform [35].

● **The fast Fourier transform (FFT)**

Discrete convolution. By analogy with the conventional convolution of two signals

$$f(t) = \int_{-\infty}^{\infty} x(\tau)\, y(t-\tau)\, d\tau$$

we may introduce a *discrete convolution* which is a signal represented by a collection of its sampled values

$$f_m = \frac{1}{N}\sum_{k=0}^{N-1} x_k y_{m-k}, \quad m = 0, 1, 2, \ldots, N-1 \qquad (15.13)$$

Let us establish the relation between the coefficients of the discrete convolution and the coefficients of the signals $x_d(t)$ and $y_d(t)$. To this end, we express the instantaneous values of the coefficients x_k and y_{m-k} as the inverse discrete Fourier transforms of the corresponding spectra:

$$x_k = \sum_{n=0}^{N-1} C_{xn}\exp(j2\pi nk/N) \qquad y_{m-k} = \sum_{l=0}^{N-1} C_{yl}\exp[j2\pi l(m-k)/N]$$

and substitute them in (15.13):

$$f_m = \frac{1}{N}\sum_{k=0}^{N-1}\left[\sum_{n=0}^{N-1} C_{xn}\exp(j2\pi nk/N)\right]\left\{\sum_{l=0}^{N-1} C_{yl} \times \exp[j2\pi l(m-k)/N]\right\}$$

On changing the order of summation, we have

$$f_m = \frac{1}{N}\sum_{n=0}^{N-1}\sum_{l=0}^{N-1} C_{xn}C_{yl}\exp(j2\pi lm/N) \times \sum_{k=0}^{N-1}\exp[j2\pi(n-l)k/N] \qquad (15.14)$$

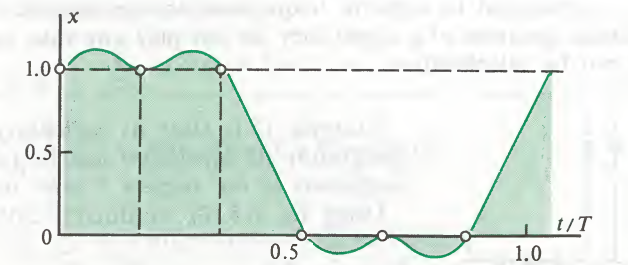


Fig. 15.5 Signal reconstituted from discrete Fourier transform coefficients

where $\varphi_i = \arg C_i$ are the phase angles of the corresponding Fourier coefficients. As an example, Fig. 15.5 shows a signal $x(t)$ reconstructed from its samples in accordance with the data given in Example 15.1. On the basis of Eq. (15.11), this signal may be defined as

$$x(t) = \frac{1}{2} + \frac{2}{3}\cos(2\pi t/T - \pi/3) + \frac{1}{6}\cos(6\pi t/T)$$

▲ Solve Problem 3

It should be especially stressed that the reconstruction of a continuous waveform by Eq. (15.11) *is not an approximate, but an exact operation* fully equivalent to obtaining the instantaneous values of a band-limited signal from its samples making up the Kotelnikov series. In some cases, however, it is preferable to use the procedure based on the discrete Fourier transform, because it leads to finite sums of harmonics, while the Kotelnikov series for a periodic signal must in principle contain an infinite number of members.

The inverse discrete Fourier transform (IDFT). The task of discrete spectral analysis may be stated in a different way. Suppose that the discrete Fourier coefficients C_n are specified in advance. Also let us set $t = k\Delta$ in Eq. (15.8) and note that the sum is taken only over a finite number of members in the series, which correspond to the harmonics contained in the spectrum of the original waveform. Hence, we obtain an equation for sampled values

$$x_k = \sum_{n=0}^{N-1} C_n \exp(j2\pi nk/N) \tag{15.12}$$

which defines the algorithm of the *inverse discrete Fourier transform (IDFT)*.

The mutually complementing equations (15.10) and (15.12) are the discrete counterparts of the usual Fourier transform pair for continuous signals.

▲ Solve Problem 4

At present, discrete spectral analysis is one of the most widely used techniques of numerical analysis of signals on a computer. As an example, Appendix 5 gives a FORTRAN subroutine for computing the discrete Fourier transform. It should be noted that with a large number of sampled values, any direct calculation of the sum in (15.10) or (15.12) involves the expenditure of a lot of computer time. To avoid this, resort is widely made to what is called as the *fast Fourier transform (FFT)*. By purely algorithmic means, it substantially reduces the amount of computations involved in finding coefficients for the discrete Fourier transform and the inverse discrete Fourier transform [35].

● The fast Fourier transform (FFT)

Discrete convolution. By analogy with the conventional convolution of two signals

$$f(t) = \int_{-\infty}^{\infty} x(\tau)\, y(t-\tau)\, d\tau$$

we may introduce a *discrete convolution* which is a signal represented by a collection of its sampled values

$$f_m = \frac{1}{N}\sum_{k=0}^{N-1} x_k y_{m-k}, \quad m = 0, 1, 2, \ldots, N-1 \tag{15.13}$$

Let us establish the relation between the coefficients of the discrete convolution and the coefficients of the signals $x_d(t)$ and $y_d(t)$. To this end, we express the instantaneous values of the coefficients x_k and y_{m-k} as the inverse discrete Fourier transforms of the corresponding spectra:

$$x_k = \sum_{n=0}^{N-1} C_{xn}\exp(j2\pi nk/N) \qquad y_{m-k} = \sum_{l=0}^{N-1} C_{yl}\exp[j2\pi l(m-k)/N]$$

and substitute them in (15.13):

$$f_m = \frac{1}{N}\sum_{k=0}^{N-1}\left[\sum_{n=0}^{N-1} C_{xn}\exp(j2\pi nk/N)\right]\left\{\sum_{l=0}^{N-1} C_{yl} \times \exp[j2\pi l(m-k)/N]\right\}$$

On changing the order of summation, we have

$$f_m = \frac{1}{N}\sum_{n=0}^{N-1}\sum_{l=0}^{N-1} C_{xn} C_{yl}\exp(j2\pi lm/N) \times \sum_{k=0}^{N-1}\exp[j2\pi(n-l)k/N] \tag{15.14}$$

It is easy to note that the inner sum

$$\sum_{k=0}^{N-1} \exp[j2\pi(n-l)k/N] = \begin{cases} N, & \text{if } n = l \\ 0, & \text{if } n \neq l \end{cases}$$

Here, the upper equality is obvious; the sum vanishes for $n \neq l$ because all terms are complex numbers of unity magnitude and with a linearly increasing argument. When they are combined vectorially, they always form a regular closed polygon on the complex plane. Taking advantage of the above result, we obtain on the basis of Eq. (15.14)

The convolution of periodic signals presented here is sometimes called cyclic convolution

$$f_m = \sum_{n=0}^{N-1} C_{xn} C_{yn} \exp(j2\pi mn/N) \tag{15.15}$$

Since Eq. (15.15) is an inverse discrete Fourier transform, we may conclude that the Fourier coefficients of the convolution are the products of the discrete Fourier coefficients of the signals being convoluted:

$$C_{fk} = C_{xk} C_{yk}, \quad k = 0, 1, \ldots, N-1 \tag{15.16}$$

Apart from establishing a useful analogy between the spectral properties of continuous and discrete signals, the above result is of importance for the theory of discrete signals and digital filters. The point is that if the number of sampled values is very large (say, several thousand), then, in finding the convolution of signals, one should first find their discrete Fourier transforms, then multiply the coefficients, and finally use Eq. (15.15) and the fast Fourier transform. This procedure is often more economical than the direct use of Eq. (15.13).

15.3 The Theory of the z-Transform

The z-transform is widely used in the analysis and synthesis of sampled-data systems. It has the same relationship to discrete (or, rather, sampled) signals as the integral Fourier and Laplace transformations bear to continuous signals. This section will set forth the basic theoretical aspects of the z-transform related to the properties of this functional transformation.

Definition of the z-transform. Let $\{x_k\} = (x_0, x_1, x_2, \ldots)$ be a numerical sequence, finite or infinite, which contains the sampled values of a signal. We place this sequence in a one-to-one correspondence with a sum of an inverse power series of the complex variable z:

$$X(z) = x_0 + x_1/z + x_2/z^2 + \ldots = \sum_{k=0}^{\infty} x_k z^{-k} \tag{15.17}$$

This sum, if it exists, is called the *z-transform* of the sequence $\{x_k\}$. The introduction of this mathematical concept is warranted because the properties of discrete sequences of numbers can conveniently be analysed by investigating their z-transforms with the usual techniques of mathematical analysis.

In mathematics, the z-transform is also called the generating function of the original sequence

On the basis of Eq. (15.17) we directly derive z-transforms for sampled signals with a finite number of sampled values. Thus the simplest sampled signal with a single sampled value, $\{x_k\} = (1, 0, 0, \ldots)$, corresponds to $X(z) = 1$. If, for example, $\{x_k\} = (1, 1, 1, 0, 0, 0, \ldots)$, then

$$X(z) = 1 + 1/z + 1/z^2 = (z^2 + z + 1)/z^2$$

▲ Solve Problem 5

Convergence of the series. If the series defined in (15.17) contains an infinite number of members with nonzero coefficients, we should inquire into the convergence of the series. From the theory of functions of a complex variable [11] the following is known. Let the coefficients of the series in question satisfy the condition

$$|x_k| < MR_0^k \tag{15.18}$$

for any $k \geqslant 0$. Here, $M > 0$ and $R_0 > 0$ are constant real numbers. Then the series (15.17) converges for any z such that $|z| > R_0$. In this region of convergence, the sum of the series is an analytic function of the variable z, having neither poles nor substantially singular points.

As an example, consider a sampled signal $\{x_k\} = (1, 1, 1, \ldots)$ formed by identical samples of value unity and serving as a model for the usual switching function. The infinite series

$$X(z) = 1 + 1/z + 1/z^2 + \ldots$$

is the sum of a geometric progression and it converges for any z inside the domain $|z| > 1$. On taking the sum of the progression, we get

$$X(z) = 1/(1 - 1/z) = z/(z-1)$$

At the boundary of analyticity, this function has only one simple pole for $z = 1$.

Acting similarly, we derive the z-transform of an infinite sampled signal of the form

$$\{x_k\} = (1, a, a^2, \ldots)$$

where a is some real number. Then

$$X(z) = \frac{1}{1 - a/z} = z/(z-a)$$

with the series converging and with the expansion having any meaning only inside the domain $|z| > a$.

▲ **Solve Problem 6**

The z-transform of continuous functions. Assuming that $\{x_k\}$ are the sampled values of the continuous function $x(t)$ at the sampling instants $t = k\Delta$, we can assign to any signal $x(t)$ its z-transform with a suitably selected sampling interval:

$$X(z) = \sum_{k=0}^{\infty} x(k\Delta) z^{-k} \tag{15.19}$$

For example, if $x(t) = \exp(\alpha t)$, then the corresponding z-transform

$$X(z) = \sum_{k=0}^{\infty} \exp(\alpha k\Delta) z^{-k} = \frac{z}{z - \exp(\alpha\Delta)}$$

is an analytic function for $|z| > \exp(\alpha\Delta)$.

The inverse z-transform. Let $X(z)$ be a function of the complex variable z, analytic inside the circle $|z| > R_0$. A distinction of the z-transform is that the function $X(z)$ possessing this property defines all of the infinite set of sampled coefficients $(x_0, x_1, x_2, \ldots)$.

To demonstrate, we multiply both sides of Eq. (15.17) by z^{m-1}:

$$z^{m-1} X(z) = x_0 z^{m-1} + x_1 z^{m-2} + \ldots + x_m z^{-1} + \ldots \tag{15.20}$$

then evaluate the integrals of the both sides of the above equality, choosing as the contour of integration a closed path wholly lying in the analyticity region and enclosing all the poles of the function $X(z)$. In doing so, we take advantage of the fundamental fact stemming from Cauchy's theorem:

The integration contour is traversed in the positive, that is, counterclockwise direction

$$\oint z^n \, dz = \begin{cases} 2\pi j, & \text{if } n = -1 \\ 0, & \text{if } n \neq -1 \end{cases}$$

Then the integrals of all terms on the right-hand side, except the mth term, will vanish, and so

$$x_m = \frac{1}{2\pi j} \oint z^{m-1} X(z) \, dz \tag{15.21}$$

● **The inverse z-transform**

Equation (15.21) is known as the *inverse z-transform.*

Example 15.2. *Given: The z-transform*

$X(z) = (z + 1)/z$

To find: The coefficients of the sampled signal answering the above function.

To begin with, we note that $X(z)$ is analytic over the entire plane, except the point $z = 0$, and so there may exist a z-transform of some sampled (discrete) signal.

▲ **Solve Problem 7**

Referring to Eq. (15.21), we find that

$$x_0 = \frac{1}{2\pi j} \oint \frac{z+1}{z^2} \, dz = 1$$

$$x_1 = \frac{1}{2\pi j} \oint \frac{z+1}{z} \, dz = 1$$

$$x_m = \frac{1}{2\pi j} \oint z^{m-2} (z - 1) \, dz = 0$$

for any $m \geqslant 2$. Thus, the sampled signal in question has the form (1, 1, 0, 0, 0, ...).

Relation to the Laplace and Fourier transforms. Let us define for $t \geqslant 0$ a sampled (discrete) signal in the form of an ideal pulse train:

$$x_d(t) = \sum_{k=0}^{\infty} x_k \delta(t - k\Delta)$$

Its Laplace transform is

$$F(p) = \sum_{k=0}^{\infty} x_k \exp(-pk\Delta) \tag{15.22}$$

which directly goes into the z-transform if we set $z = \exp(p\Delta)$. If, on the other hand, we put $z = \exp(j\omega\Delta)$, the expression

$$S(\omega) = \sum_{k=0}^{\infty} x_k \exp(-j\omega k\Delta) \tag{15.23}$$

will be the Fourier transform of the pulse train.

The relation we have just established plays an important role and permits us to draw certain analogies between the properties of continuous and discrete (sampled) signals.

The most important properties of the z-transform.

The properties listed here are directly analogous to those of the Fourier and Laplace transforms of analog signals

1. *Linearity.* If $\{x_k\}$ and $\{y_k\}$ are two numeric sequences representing some sampled signals whose z-transforms $X(z)$ and $Y(z)$ are known, then the signal $\{u_k\} = \{\alpha x_k + \beta y_k\}$ corresponds to the transform $U(z) = \alpha X(z) + \beta Y(z)$ for any constants α and β. This is proved by substituting the sum in Eq. (15.17).

2. *The z-transform of a time-shifted signal.* Consider a sampled signal $\{y_k\}$ obtained by shifting $\{x_k\}$ backward by one position (one sampling interval) such that $y_k = x_{k-1}$. By directly evaluating the z-transform, we get

$$Y(z) = \sum_{k=0}^{\infty} x_{k-1} z^{-k} = z^{-1} \sum_{n=0}^{\infty} x_n z^{-n} = z^{-1} X(z) \tag{15.24}$$

The unit delay operator

It is to be noted that z^{-1} is the unit delay operator (a delay by one sampling interval) in the z-domain.

3. *The z-transform of a convolution.* Let $x(t)$ and $y(t)$ be continuous signals for which the convolution function is

$$f(t) = \int_{-\infty}^{\infty} x(\tau)\, y(t-\tau)\, \mathrm{d}\tau = \int_{-\infty}^{\infty} y(\tau)\, x(t-\tau)\, \mathrm{d}\tau \qquad (15.25)$$

By analogy with Eq. (15.25), for sampled signals we introduce a discrete convolution $\{f_k\}$ which is a series of numbers the common term of which is

In contrast to the circular convolution, it is sometimes called the linear convolution

$$f_m = \sum_{k=0}^{\infty} x_k y_{m-k} = \sum_{k=0}^{\infty} y_k x_{m-k}, \quad m = 0, 1, 2, \ldots \qquad (15.26)$$

Let us take the z-transform of a discrete convolution:

$$F(z) = \sum_{m=0}^{\infty} \sum_{k=0}^{\infty} x_k y_{m-k} z^{-m} = \sum_{m=0}^{\infty} \sum_{k=0}^{\infty} x_k z^{-k} y_{m-k} z^{-(m-k)}$$

$$= \sum_{k=0}^{\infty} x_k z^{-k} \sum_{n=0}^{\infty} y_n z^{-n} = X(z)\, Y(z) \qquad (15.27)$$

Properties of the discrete convolution

Thus, the convolution of two sampled signals corresponds to the product of their z-transforms.

15.4 Digital Filters

At present, communication signals are frequently processed with the aid of microcomputers and microcomputer systems. This section deals with the simplest and best known class of systems for digital signal processing known as *linear stationary digital filters*. Performing, similarly to analog circuits, the operation of frequency filtering, digital filters offer a number of important advantages. Among other things, this includes the high parameter stability and the choice of an extremely wide variety of shapes for the amplitude and phase responses. Digital filters need no alignment and tuning, and can readily be implemented in the form of computer algorithms and programs.

The principle of digital signal processing. A basic block diagram for digital signal processing is shown in Fig. 15.6.

The analog-to-digital converter

A continuous input signal $x(t)$ is fed to an *analog-to-digital converter, ADC*, controlled by sync pulses from a special generator which maintains the desired sampling rate. Each time a sync pulse is applied, the ADC delivers at its output a signal representing the instantaneous value of the input wave in the form of a binary number with a fixed number of digit (bit) positions. Depending on

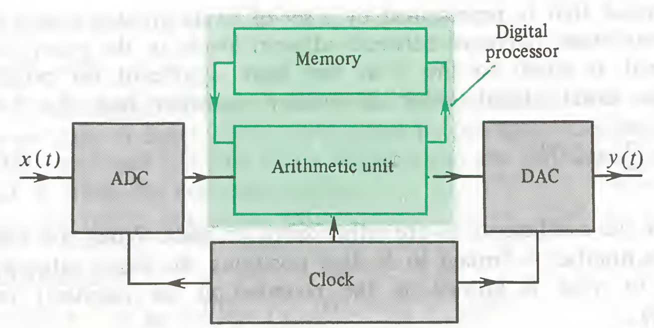


Fig. 15.6 Block diagram of a set-up for digital processing of continuous signals

the particular arrangement of the digital filter, this number may correspond to a train of short pulses (when the transmission is in a sequential code) or to a set of voltage levels at the signal lines of the individual bit positions (when the transmission is in a parallel code). Thus transformed, the signal is fed to the main unit of the system, called the *digital signal processor* consisting of an arithmetic unit and a memory (or storage) unit. The arithmetic unit performs on numbers a range of operations such as multiplication, addition, and a time shift through a specified number of sampling intervals. The memory unit may store a number of previous sampled values of the input and output signals, which may be utilized in the course of signal processing.

The digital signal processor

The digital processor handles the incoming numbers as prescribed by a filtering algorithm and produces at its output a sequence of binary numbers representing the output signal. If the end user needs data in analog form, the number sequence is additionally passed through a *digital-to-analog converter, DAC*. However, a digital filter may contain no DAC, if the output signal is only to be subjected to digital processing.

The digital-to-analog converter

The performance of a digital filter is judged in terms of its speed which depends on the rate of transients in the circuit components and on the complexity of the filtering algorithm.

In the early 70s the frequency limit for the signals processed by digital filters was a few kilohertz. Advances in present-day microelectronics are continuously pushing back this limit. In recent time, digital signal filtering has been given a new momentum in its progress by the advent of relatively inexpensive and reliable microprocessors and memory units based on large-scale and very-large-scale integrated (LSI and VLSI) circuits.

Signal quantization in digital filters. A distinction of any digital device is that signals are represented as numeric sequences with a limited number of digit positions. Therefore, the sampled signal is

quantized, that is, represented by a set of levels in such a way that the minimum difference between adjacent levels, or the *quantization interval*, is equal to the 1 in the least significant bit position.

The exact signal value in binary notation has the form:

$$\tilde{x} = \sum_{m=0}^{\infty} \alpha_n 2^{-n} \tag{15.28}$$

where the coefficients α_n are either zeros or ones. When the length of the number is limited to N digit positions, the exact value gives way to what is known as the rounded-off (or *machine*) value

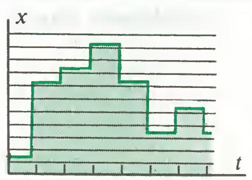
The machine representation of a number in a digital filter

$$x = \sum_{m=0}^{N-1} \alpha_n 2^{-n} + \beta_N 2^{-N} \tag{15.29}$$

where the coefficient β_N is equal to either α_N or $\alpha_N + 1$, according as the $(N+1)$st bit position contains a zero or a one.

Quantized signals

In communication engineering, sampled signals represented by a countable set of levels or amplitudes are called *amplitude-quantized* or, simply, *quantized signals*. Signal quantization leads to a difference between the input signal and the quantized output. This difference is called the *quantization noise*. A straightforward path to minimize the quantization noise is to use multibit binary numbers. However, this inevitably reduces the speed of the digital filter because it takes more time to process multibit numbers. Therefore, practical microprocessors employed in sampled-data processing and sampled-data control systems ordinarily use binary numbers with 4 to 16 bit positions.



It may be added that quantization is also used in continuous-signal techniques, especially in various types of noise-immune pulse-modulated systems.

The algorithm of linear digital filtering. The mathematical theory of digital filters carries over to the case of discrete signals all the basic considerations known from the theory of linear systems used to transform continuous signals.

As will be recalled, a linear stationary system transforms a continuous input signal $x(t)$ in such a way that its output delivers a wave $y(t)$ equal to the convolution of $x(t)$ and the impulse response $h(t)$:

$$y(t) = \int_{-\infty}^{\infty} x(\tau) h(t-\tau)\, d\tau \tag{15.30}$$

A digital filter is a system (which may be a physical device or a computer program) which transforms a series, $\{x_k\}$, of samples of the input signal into a series, $\{y_k\}$, of samples of the output signal:

$$(x_0, x_1, x_2, \ldots) \Rightarrow (y_0, y_1, y_2, \ldots) \tag{15.31}$$

or, in compact form,

$$\{x_k\} \Rightarrow \{y_k\}$$

A linear digital filter is characterized by the fact that the sum of any number of input signals multiplied by arbitrary coefficients is transformed into the sum of its responses to the individual terms. That is, from the correspondences

$$\{x_k^{(1)}\} \Rightarrow \{y_k^{(1)}\}, \ldots, \{x_k^{(N)}\} \Rightarrow \{y_k^{(N)}\}$$

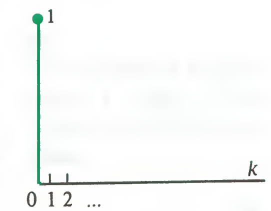
it follows that

$$\alpha_1 \{x_k^{(1)}\} + \ldots + \alpha_N \{x_k^{(N)}\} \Rightarrow \{\alpha_1 y_k^{(1)} + \ldots + \alpha_N y_k^{(N)}\} \tag{15.32}$$

for any coefficients $\alpha_1, \alpha_2, \ldots, \alpha_N$.

The impulse response of a digital filter

So that Eq. (15.30) can be generalized to the discrete case, we introduce the concept of the *impulse response* for the digital filter. By definition, this is a discrete signal $\{h_k\}$ which is the response of the digital filter to a "unit impulse" (1, 0, 0, 0, ...):

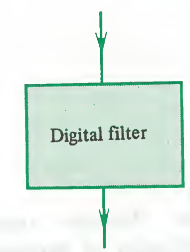


$$(1, 0, 0, 0, \ldots) \Rightarrow (h_0, h_1, h_2, \ldots) \tag{15.33}$$

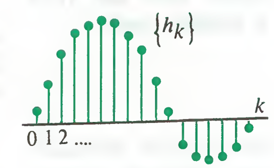
A linear digital filter is stationary if the shift of the input unit impulse by any number of sampling intervals causes the impulse response to be shifted in a similar way, without any change in shape. For example:

$$(0, 1, 0, 0, \ldots) \Rightarrow (0, h_0, h_1, h_2, \ldots) \tag{15.34}$$

$$(0, 0, 1, 0, \ldots) \Rightarrow (0, 0, h_0, h_1, h_2, \ldots)$$



Let us see how the properties of linearity and stationarity lead to the most common algorithm for the digital filter in question. Let $\{x_k\} = (x_0, x_1, x_2, \ldots)$ be a signal applied to the input of the digital filter. The impulse response of the filter is known in advance. Using Eqs. (15.32) and (15.34), we may directly write the mth sampled value of the output signal $\{y_k\}$:



$$y_m = x_0 h_m + x_1 h_{m-1} + \ldots + x_m h_0 = \sum_{k=0}^{m} x_k h_{m-k} \tag{15.35}$$

Equation (15.35), which plays a leading role in the theory of linear digital filtering, tells us that the output series is the discrete convolution of the output signal and the impulse response of the filter. The meaning of the equation is simple and easy-to-visualize:

The physical realizability of a digital filter

At each sampling instant, a digital filter performs the operation of the weighted summation of all previous values of the input signal, with the samples of the impulse response playing the role of the series of weighting coefficients. In other words, a digital filter has some sort of "memory" for the past inputs.

Only physically realizable digital filters are of practical interest. For a physically realizable digital filter, its impulse response cannot be nonzero at the sampling instants preceding the instant when the input pulse is applied. Therefore, for physically realizable systems the coefficients $h_{-1}, h_{-2}, \ldots$ vanish, and the sum in (15.35) may be taken solely over all positive values of the index k:

$$y_m = \sum_{k=0}^{\infty} x_k h_{m-k}, \quad m = 0, 1, 2, \ldots \tag{15.36}$$

A harmonic series or sequence

Discrete harmonic sequences. As will be recalled, complex signals of the form $x(t) = A\exp[j(\omega t + \varphi)]$, representing harmonic waves play a special role in the theory of linear systems. When such a signal is sampled, the result is a *harmonic series* or *sequence*

$$\{x_k\} = \{A\,e^{j(\omega k\Delta + \varphi)}\} \tag{15.37}$$

such that

$$\operatorname{Re}\{x_k\} = \{A\cos(\omega k\Delta + \varphi)\} \tag{15.38}$$

The discrete representation of harmonic signal does not yield a unique result

It should be borne in mind that the sequences defined in (15.37) and (15.38) do not represent sampled harmonic signals in a unique manner. Indeed, these sequences will not change if we replace the frequency ω with $\omega + 2\pi n/\Delta = \omega + n\omega_s$, where n is any integer and ω_s is the angular sampling frequency or rate. Therefore, we cannot in principle tell from one another two sampled harmonic waves if they differ in frequency by a whole-number multiple of the sampling rate.

The frequency response of a digital filter. Suppose that the input of a linear stationary digital filter accepts a harmonic sequence $\{x_k\}$ such as defined in (15.37), which is unbounded in time, that is, with the index k capable of taking values $0, \pm 1, \pm 2, \ldots$. In order to find the output signal of the filter, $\{y_k\}$, we take advantage of the convolution in (15.35) and determine the mth sample at the output:

$$y_m = \sum_{k=-\infty}^{m} x_k h_{m-k} = A\exp(j\varphi) \sum_{k=-\infty}^{m} \exp(j\omega k\Delta)\, h_{m-k}$$

On carrying out identity transformations, we get

$$y_m = A\,e^{j(\omega m\Delta + \varphi)} \sum_{k=-\infty}^{m} e^{j\omega(k-m)\Delta} h_{m-k}$$

Let us introduce a new summation index $n = m - k$. Then

$$y_m = A\,e^{j(\omega m\Delta + \varphi)} \sum_{n=0}^{\infty} e^{-j\omega n\Delta} h_n \tag{15.39}$$

The structure of the output signal of a digital filter

As follows from Eq. (15.39), the output signal has the structure of a discrete harmonic sequence of the same frequency as the input signal. The output samples are derived from the input samples by multiplying the latter by the complex number

$$K(j\omega) = \sum_{n=0}^{\infty} \exp(-j\omega n\Delta)\, h_n \tag{15.40}$$

called the *frequency response* of the digital filter. It depends on the frequency ω of the input signal, the sampling interval Δ, and the set of coefficients, $\{h_n\}$, of the impulse response of the digital filter.

The frequency response of a digital filter shows a periodic behaviour

From Eq. (15.40) we may draw the following important conclusions:

1. The frequency response of a digital filter is a periodic frequency function with a period equal to the sampling rate $\omega_s = 2\pi/\Delta$.

2. The function $K(j\omega)$ may be treated as the Fourier transform of the impulse response of a digital filter, represented by a series of delta impulses:

$$h_d(t) = h_0\,\delta(t) + h_1\,\delta(t-\Delta) + \ldots$$

(Compare with Eq. (15.23).)

The system function of a digital filter. The frequency response of a digital filter can conveniently be found using the techniques of the z-transformation.

Given the three sequences of discrete signals $\{x_k\}$, $\{y_k\}$ and $\{h_k\}$, their z-transforms will be $X(z)$, $Y(z)$ and $H(z)$, respectively. The output signal of a digital filter, $\{y_k\}$, is the convolution of the input signal and the impulse response, so [see Eqs. (15.27) and (15.35)], the function corresponding to the output signal is

$$Y(z) = H(z)\,X(z) \tag{15.41}$$

Relation between the system function and the impulse response of a digital filter

The ratio of the z-transform of the output signal to the z-transform of the input signal is called the *system function* of a stationary linear digital filter. From Eq. (15.41) we find that the system function of such a filter, defined by

$$H(z) = Y(z)/X(z) = \sum_{k=0}^{\infty} h_k z^{-k} \tag{15.42}$$

is the z-transform of its impulse response.

Comparing Eqs. (15.40) and (15.42), we can draw the following conclusion: The frequency response of a digital filter can be deduced by inserting $z = \exp(j\omega\Delta)$ in its system function.

Example 15.3. *Consider a digital filter whose impulse response consists of two nonzero samples:* $\{h_k\} = (1, -1, 0, 0, 0, \dots)$ *and find its frequency response* $K(j\omega)$.

Here, the system function is

$$H(z) = 1 - z^{-1}$$

Hence,

$$K(j\omega) = 1 - \exp(-j\omega\Delta) = (1 - \cos\omega\Delta) + j\sin\omega\Delta$$

$$= |K(j\omega)|\exp[j\varphi_K(\omega)]$$

The amplitude response of the filter takes the form

$$|K(j\omega)| = \sqrt{(1-\cos\omega\Delta)^2 + \sin^2\omega\Delta} = 2\left|\sin\frac{\omega\Delta}{2}\right|$$

and its phase response is

$$\varphi_K(\omega) = \arctan\frac{\sin\omega\Delta}{1-\cos\omega\Delta}$$

The amplitude and phase characteristics of the filter are periodic functions of frequency, but practically they are meaningful only in the interval from 0 to $\omega = \pi/\Delta$. At the upper frequency limit of this interval, two samples correspond to each period of the sampled harmonic signal. By the Kotelnikov sampling theorem, this is the highest frequency of the signal that can be uniquely recovered from its samples.

▲ **Work Problem 8**

It is to be noted that if the input to such a filter is a harmonic signal at a frequency substantially below the sampling rate such that $\omega\Delta \ll 1$, then

$$K(j\omega) \approx \left[1 - 1 + \frac{(\omega\Delta)^2}{2} - \dots\right] + j\left[\omega\Delta - \frac{(\omega\Delta)^3}{6} + \dots\right]$$

$$\approx j\omega\Delta$$

Therefore, the system in question performs the operation of approximate differentiation with respect to the slow input signals.

15.5 Implementation of Digital Filtering Algorithms

In order to form the output signal corresponding to the ith sampling instant, physically realizable filters operating in real time may use the following initial specifications: (a) the value of the

input signal at the ith sampling instant and also a number of "past" input samples $x_{i-1}, x_{i-2}, \dots, x_{i-m}$; (b) several past samples of the output signal, $y_{i-1}, y_{i-2}, \dots, y_{i-n}$. The integers m and n define the *order* of a digital filter.

● **The order of a digital filter**

According to the manner in which information about the past states of the system is utilized, digital filters can be classed into several types.

Transversal digital filters. This term refers to digital systems which operate by the following algorithm:

$$y_i = a_0 x_i + a_1 x_{i-1} + a_2 x_{i-2} + \dots + a_m x_{i-m} \tag{15.43}$$

where $a_0, a_1, \dots, a_m$ are a series of coefficients. The number m is the order of the transversal filter. As is seen from Eq. (15.43), a transversal filter carries out the weighted summation of the previous samples of the input signal and does not use any previous samples of the output signal. By taking the z-transforms of both sides of Eq. (15.43), we see that

$$Y(z) = (a_0 + a_1 z^{-1} + a_2 z^{-2} + \dots + a_m z^{-m}) X(z)$$

Hence, the system function of the filter

$$H(z) = a_0 + a_1 z^{-1} + a_2 z^{-2} + \dots + a_m z^{-m}$$

$$= \frac{a_0 z^m + a_1 z^{m-1} + a_2 z^{m-2} + \dots + a_m}{z^m} \tag{15.44}$$

is a rational function of z, which has m-tiple pole for $z = 0$ and m zeros whose coordinates are defined by the coefficients of the filter.

■ **The form of the system function of a transversal filter**

The algorithm by which a transversal digital filter operates is illustrated by the block diagram of Fig. 15.7. The basic building blocks of the filter are the delay units, z^{-1}, which delay the samples for one sampling interval, and also the scalers which multiply by the corresponding coefficients in digital form. From the scalers, the signals are routed to an adder where they are combined to yield the sampled value of the signal.

From the structure of the block diagram, it is clear why this filter is called transversal.

In a digital processor the samples are delayed for one bit by a shift register

Program (software) implementation of a transversal filter. It should be borne in mind that the block diagram appearing in Fig. 15.7 is not a kind of schematic diagram for some electric network, but only a graphical representation of the algorithm for signal processing. As an example, consider a fragment of a FORTRAN routine implementing transversal digital filtering.

Let the internal computer memory contain two one-dimensional arrays of M locations each. One array, X, contains the values of the input signal; the other array, A, holds the coefficients of the filter. The data in the arrays are disposed in the following manner.

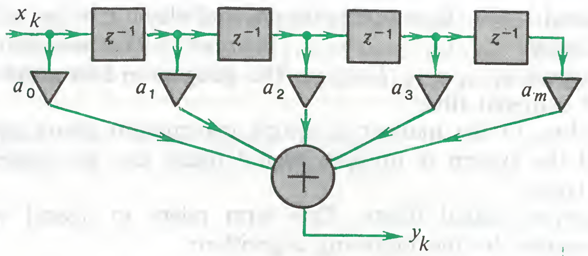


Fig. 15.7 Block diagram of a transversal digital filter

Array A holds constants, whereas the contents of the locations in array X change each time a new sampled value of the input signal is received. Suppose that array X is filled full with the previous sampled values of the input series and consider the situation arising when there arrives the next sampled value which the program has named the variable S. This sampled value must be stored in location 1, and this can happen only after the previous record has been shifted one position to the right, that is, backward.

The elements of array X thus formed are multiplied one by one by the elements of array A, and the result is stored in location Y which accumulates the sampled value of the output signal. The routine for transversal digital filtering is given below:

It is assumed that the input samples and the filter coefficients are represented by real numbers

```
. . . . . . . . . . . . . . . . . . . . . .
  DO1 K = 1, M - 1
1 X(M - K + 1) = X(M - K)
  X(1) = S
C ARRAY X IS FORMED
  Y = 0
  DO2 K = 1, M
2 Y = Y + X(K) * A(K)
```

The form of the impulse response of a transversal filter

The impulse response. Let us go back to Eq. (15.44) and find the impulse response for a transversal digital filter by taking the inverse z-transform. It is easy to see that each term of the system function $H(z)$ is responsible for a contribution which is equal to the corresponding coefficient a_j shifted j bits backward. Thus, here

$$\{h_k\} = (a_0,\ a_1,\ a_2,\ \ldots,\ a_m) \tag{15.45}$$

This result can be obtained directly by inspection of the block diagram of the filter in Fig. 15.7, assuming that it is driven by a unit pulse (1, 0, 0, 0, ...).

It is important to note that the impulse response of a transversal filter contains a finite number of members.

The frequency response. By a change of variable, $z = \exp(j\omega\Delta)$ in Eq. (15.44), we get

$$K(j\omega) = a_0 + a_1 e^{-j\omega\Delta} + a_2 e^{-j2\omega\Delta} + \ldots + a_m e^{-jm\omega\Delta} \tag{15.46}$$

With any specified sampling interval Δ, we can obtain a fairly wide variety of amplitude and phase characteristics by choosing appropriate weighting coefficients for the filter.

Example 15.4. *Investigate a second-order transversal digital filter which takes the average of the present value of the input signal and of its two previous samples:*

This type of filter is said to perform smoothing by trees

$$y_i = \frac{1}{3}(x_i + x_{i-1} + x_{i-2}) \tag{15.47}$$

The system function of the filter is

$$H(z) = \frac{1}{3}(1 + z^{-1} + z^{-2})$$

Hence, the frequency response of the filter is

$$K(j\omega) = \tfrac{1}{3}(1 + e^{-j\omega\Delta} + e^{-j2\omega\Delta})$$

$$= \frac{1}{3}[(1 + \cos\omega\Delta + \cos 2\omega\Delta) - j(\sin\omega\Delta + \sin 2\omega\Delta)]$$

Simple manipulations lead to the following expression for the amplitude response:

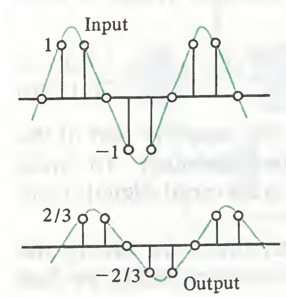
Solve Problem 9

$$|K(j\omega)| = \frac{1}{3}\sqrt{3 + 4\cos\omega\Delta + 2\cos 2\omega\Delta} \tag{15.48}$$

and for the phase response of the system:

$$\varphi_K(\omega) = -\arctan\frac{\sin\omega\Delta + \sin 2\omega\Delta}{1 + \cos\omega\Delta + \cos 2\omega\Delta} = -\omega\Delta \tag{15.49}$$

Their plots appear in Fig. 15.8 where the phase angle, $\omega\Delta$, of the sampling interval for a given value of frequency is laid off as abscissa.



Suppose that $\omega\Delta = 60°$, which means that six samples are taken over each period of the harmonic input wave. Then the input sequence will have, say, the following pattern

$$\ldots,\ 0,\ 1,\ 1,\ 0,\ -1,\ -1,\ 0,\ 1,\ 1,\ \ldots$$

(the absolute values of samples are immaterial, because the filter is linear). Using the algorithm in (15.47), we find the output sequence:

$$\ldots,\ \frac{2}{3},\ \frac{2}{3},\ 0,\ -\frac{2}{3},\ -\frac{2}{3},\ 0,\ \ldots$$

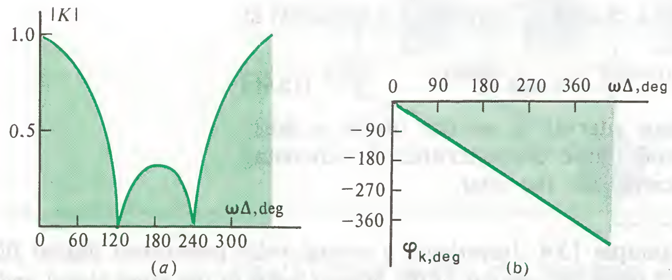


Fig. 15.8 Characteristics of the transversal digital filter examined in Example 15.4: (*a*) amplitude response; (*b*) phase response

As is seen, it corresponds to a harmonic output signal of the same frequency as the input signal, but with an amplitude which is $2/3 = 0.66$ the amplitude of the input wave, and with the initial phase shifted by 60° lagging.

If $\omega\Delta < 120°$, the filter could act as a low-pass filter since it smoothes the input sequence. However, its frequency response is periodic and nonmonotonic. If the original analog signal has not been subjected to preliminary frequency filtering and contains components for which $\omega\Delta > 180°$ (so that the requirement of the Kotelnikov sampling theorem is not satisfied), they will not be attenuated by the digital filter in question. In fact, from the samples of these high-frequency components the digital-to-analog converter would reconstruct some low-frequency wave not present in the input signal. This undesirable event (the effect of superposition or masking by the high-frequency components of the spectrum) is in principle inherent in any digital system and makes mandatory the preliminary processing of signals which are to be subjected to digital filtering.

● The effect of superposition

Recursive digital filters. A distinction of this type of digital filter is that the *i*th sample is formed using the previous values of both the input and output sequences:

$$y_i = a_0 x_i + a_1 x_{i-1} + \ldots + a_m x_{-m} + b_1 y_{i-1} + b_2 y_{i-2} + \ldots + b_n y_{i-n} \qquad (15.50)$$

where the coefficients $b_1, b_2, \ldots, b_n$ defining the recursive part of the filtering algorithm are not equal to zero simultaneously. To stress the difference in structure, it is usual to call transversal digital filters nonrecursive.

The system function of a recursive digital filter. On taking the *z*-transforms of both parts of the recursion relation (15.50), we find

that the system function

$$H(z) = Y(z)/X(z) = \frac{a_0 + a_1 z^{-1} + \ldots + a_n z^{-m}}{1 - b_1 z^{-1} - \ldots - b_n z^{-n}}$$

$$= \frac{a_0 z^n + a_1 z^{n-1} + \ldots + a_m z^{n-m}}{z^n - b_1 z^{n-1} - \ldots - b_n} \qquad (15.51)$$

describing the frequency behaviour of a recursive digital filter has n poles in the z-plane. If the coefficients of the recursive part are real, the poles either lie on the real axis or form complex-conjugate pairs.

Recursion is a mathematical technique consisting in that the data obtained at the previous steps are cyclically referred to

■ The form of the system function of a recursive filter

The block diagram of a recursive digital filter. In block-diagram form, the algorithm answering Eq. (15.50) appears in Fig. 15.9. As is seen, the top part of the block diagram corresponds to the

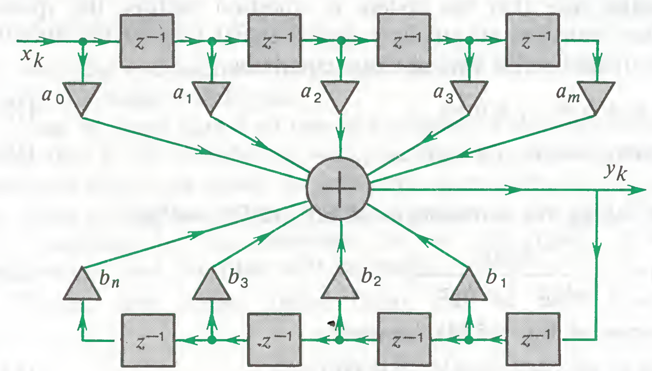


Fig. 15.9 Block diagram of a recursive digital filter

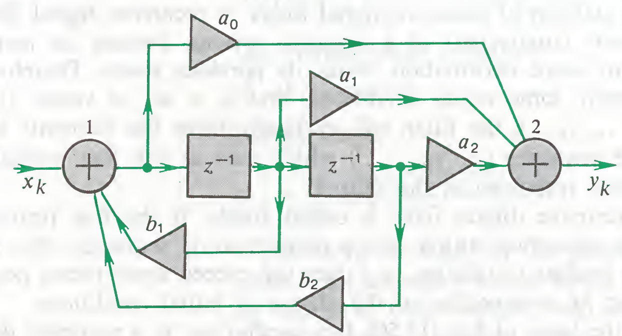


Fig. 15.10 Canonic 2nd-order recursive filter

transversal (nonrecursive) part of the filtering algorithm. In the general case, its implementation calls for $m+1$ scaling (multiplication) blocks and m memory cells to store the input samples.

The recursive part of the algorithm provides for the use of n consecutive values of the output signal, which are shifted from location to location during the operation of the filter.

The canonic form of a digital filter

A disadvantage of the above realization principle is that it calls for a large number of memory cells, separately for the recursive and the nonrecursive parts. A far better choice is offered by canonic recursive digital filters which use the least possible number of memory cells equal to the largest of the numbers m and n. As an example, Fig. 15.10 shows the block diagram of a 2nd-order canonical filter for which the system function has the form

$$H(z)=\frac{a_0+a_1z^{-1}+a_2z^{-2}}{1-b_1z^{-1}-b_2z^{-2}} \tag{15.52}$$

To make sure that the system in question realizes the specified function, consider an auxiliary digital signal $\{w_k\}$ at the output of adder *1*, and write two obvious equations:

$$w_k=x_k+b_1w_{k-1}+b_2w_{k-2} \tag{15.53}$$

$$y_k=a_0w_k+a_1w_{k-1}+a_2w_{k-2} \tag{15.54}$$

On taking the z-transform of Eq. (15.53), we get

$$W(z)=\frac{X(z)}{1-b_1z^{-1}-b_2z^{-2}} \tag{15.55}$$

By virtue of Eq. (15.54), however,

$$Y(z)=(a_0+a_1z^{-1}+a_2z^{-2})W(z) \tag{15.56}$$

On combining (15.55) and (15.56), we arrive at the specified system function in (15.52).

The stability of a digital filter

The stability of recursive digital filters. A recursive digital filter is a discrete counterpart of a dynamic system, because its memory elements store information about its previous states. Therefore, if we specify some initial conditions, that is, a set of values $(y_{i-1}, y_{i-2}, \dots, y_{i-n})$, the filter will cyclically form the elements of an infinite sequence $(y_i, y_{i+1}, \dots)$ which acts as the free oscillations (transient response) at the output.

A recursive digital filter is called stable, if the free (transient) process occurring within it is a nonincreasing sequence, that is, if, with n tending to infinity, $\{y_n\}$ does not exceed some preset positive number M, irrespective of the choice of initial conditions.

On the basis of Eq. (15.50), free oscillations in a recursive digital filter must be the solution of the following linear difference equation:

$$y_i=b_1y_{i-1}+b_2y_{i-2}+\dots+b_ny_{i-n} \tag{15.57}$$

Transversal digital filters are not dynamic systems, and they are stable with any choice of coefficients

By analogy with linear differential equations, we will seek the solution of Eq. (15.57) in the form of an exponential function

$$y_i=\alpha^i \tag{15.58}$$

in which α is as yet unknown. On substituting (15.58) into (15.57) and cancelling out the common factor, we find that α must be a root of the characteristic equation

$$\alpha^n-b_1\alpha^{n-1}-b_2\alpha^{n-2}-\dots-b_n=0 \tag{15.59}$$

which, on the basis of (15.51), fully checks with the equation satisfied by the poles of the system function of a recursive digital filter.

Suppose we have found all the roots $\alpha_1, \alpha_2, \dots, \alpha_n$ of Eq. (15.59). The general solution of the difference equation (15.57) has the form

$$y_i=A_1\alpha_1^i+A_2\alpha_2^i+\dots+A_n\alpha_n^i \tag{15.60}$$

in which the coefficients $A_1, A_2, \dots, A_n$ must be matched so as to satisfy the initial conditions.

The stability criterion for a recursive digital filter

It can be noted that if no one of the poles of the system function $H(z)$, that is, the numbers $z_1=\alpha_1, \dots, z_n=\alpha_n$, exceeds unity and therefore they all lie inside the unit circle with centre at $z=0$, then, by virtue of (15.60), any free (transient) process in a digital filter will be described by the members of a converging geometric progression, and the filter will be stable.

Clearly, only stable digital filters can be used practically.

Example 15.5. *Investigate the stability of a 2nd-order recursive filter whose system function is*

$$H(z)=\frac{a_0}{1-b_1z^{-1}-b_2z^{-2}}$$

The characteristic equation

$$z^2-b_1z-b_2=0$$

has the roots

$$z_{1,2}=\frac{b_1}{2}\pm\sqrt{(b_1/2)^2+b_2}$$

The curve described by the equation $b_1^2+4b_2=0$ on the (b_1, b_2) plane is the boundary above which the poles of the system function are real, and below which they are complex conjugate.

For complex conjugate poles,

$$|z_{1,2}|^2=-b_2$$

Therefore, one of the stability boundaries is the straight line $b_2 = -1$. If we consider real poles for $b_1 > 0$, the condition of stability will have the form

$$b_1/2 + \sqrt{(b_1/2)^2 + b_2} < 1$$

or

$$\sqrt{(b_1/2)^2 + b_2} < 1 - b_1/2$$

By squaring both sides of the above inequality, we can see that the stability region is bounded by the straight line $b_2 = 1 - b_1$. The case when $b_1 < 0$ can be analysed similarly.

▲ **Solve Problem 10**

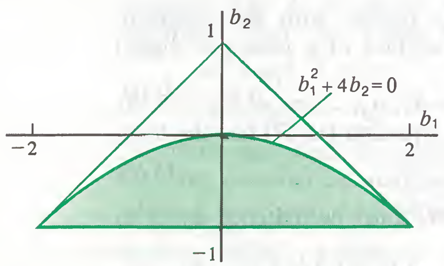


Fig. 15.11 Stability region for a 2nd-order recursive filter. (The filter poles are complex conjugate inside the coloured area.)

To sum up, the recursive filter in question is stable if the values of the coefficients b_1 and b_2 are located inside the triangle shown in Fig. 15.11.

The impulse response of a recursive digital filter. A distinction of a recursive digital filter is that it contains feedback loops. In consequence, the impulse response of a recursive digital filter is described by an unbounded time sequence. We will demonstrate the statement by reference to an elementary 1st-order filter whose system function is

$$H(z) = \frac{a}{1 - bz^{-1}} = az/(z - b)$$

The filter is stable, if $|b| < 1$.

▲ **Solve Problem 11**

As will be recalled (see Eq. (15.42)), the impulse response is found by taking the inverse z-transform of the system function. Using Eq. (15.21), we find the mth member in the sequence $\{h_k\}$:

$$h_m = \frac{1}{2\pi j}\oint \frac{az^m\,dz}{z - b} \tag{15.61}$$

The integration is carried out around the unit circle inside which the pole $z = b$ is located.

Since, as can be readily seen, the residue of the integrand at the pole is equal to ab^m, the sought impulse response of the filter is a decreasing geometric progression

$$\{h_k\} = (a,\ ab,\ ab^2,\ \ldots) \tag{15.62}$$

15.6 Synthesis of Linear Digital Filters

Currently, special emphasis in the synthesis of digital filter structures is placed on the techniques and procedures assuring some desired properties which may be a particular shape of the impulse or frequency response [35]. In the pages that follow we will be mainly concerned with synthesis procedures which substantially draw on the properties of the analog circuits that serve as prototypes for the digital filters being synthesized.

Ordinarily, a digital filter is synthesized with the aid of a computer. The number of significant digits in the final results must be sufficient for the desired accuracy to be achieved. Obviously, the filter must be stable in any case.

The invariant impulse response method. This is the simplest method for digital filter synthesis. It is based on the assumption that the impulse response of the filter being synthesized must be the same as will be produced by digitizing the impulse response of the corresponding analog prototype. Bearing in mind that we are concerned with the synthesis of physically realizable systems for which the impulse response is zero at $t < 0$, we obtain the following expression for the impulse response of a digital filter

$$\{h_k\} = (h(0),\ h(\Delta),\ h(2\Delta),\ \ldots) \tag{15.63}$$

■ **The principle of similarity between the impulse responses of analog and digital filters**

The number of samples in the impulse response sequence of a digital filter may be finite or infinite. Accordingly the structure of the filter will be different: An impulse response with a finite number of samples corresponds to a transversal filter; an impulse response with an infinite number of samples corresponds to a recursive filter.

The relation between the coefficients of the impulse response and the structure of a digital filter is especially simple in the case of a transversal filter. In the general case, the search for the desired structure is begun by taking the z-transform of the sequence in (15.63). Once the system function $H(z)$ of the filter is obtained, the next step is to compare it with the general expression (15.51) and to determine the coefficients of the transversal and recursive parts of the filter.

The degree of approximation between the frequency characteristic of the digital filter being synthesized and that of the analog prototype depends on the choice of the sampling interval Δ. If necessary, the frequency response of the digital filter can be found by a change of variable, $z = \exp(j\omega\Delta)$, in the system function $H(z)$,

and by comparing the result thus obtained with the known frequency response of the analog network.

Example 15.6. *Go through the synthesis of a transversal digital filter similar to a 1st-order dynamic system (say, an integrating RC-network) whose impulse response has the form*

$$h(t) = \begin{cases} 0, & t < 0 \\ \exp(-t/\tau), & t > 0 \end{cases} \tag{15.64}$$

(The amplitude coefficient in the impulse response is set equal to unity, as it is of minor importance to the synthesis procedure.)

Suppose that the impulse response is approximated by a sequence of three equidistant samples:

$$\{h_k\} = (1, e^{-\Delta/\tau}, e^{-2\Delta/\tau}) \tag{15.65}$$

The transversal digital filter having this form of impulse response is described by the following difference equation:

$$y_k = x_k + e^{-\Delta/\tau} x_{k-1} + e^{-2\Delta/\tau} x_{k-2} \tag{15.66}$$

By taking the z-transform of the sequence in (15.65), we obtain the system function of the digital filter:

$$H(z) = 1 + e^{-\Delta/\tau} z^{-1} + e^{-2\Delta/\tau} z^{-2} \tag{15.67}$$

Hence, the frequency response is

$$K(j\omega) = 1 + e^{-\Delta/\tau} e^{-j\omega\Delta} + e^{-2\Delta/\tau} e^{-j2\omega\Delta} \tag{15.68}$$

Example 15.7. *Consider the case where the impulse response* (15.64) *of an analog network is approximated by the infinite discrete sequence*

$$\{h_k\} = (1, e^{-\Delta/\tau}, e^{-2\Delta/\tau}, \ldots) \tag{15.69}$$

By taking the z-transform of the impulse response in (15.69), we obtain the system function

$$H(z) = 1 + e^{-\Delta/\tau} z^{-1} + e^{-2\Delta/\tau} z^{-2} + \ldots$$

$$= 1/(1 - e^{-\Delta/\tau} z^{-1}) \tag{15.70}$$

This system function defines a 1st-order recursive digital filter which contains an adder, a scaler, and a delay unit.

The frequency response of the synthesized filter is

$$K(j\omega) = \frac{1}{1 - e^{-\Delta/\tau} e^{-j\omega\Delta}} \tag{15.71}$$

Comparison of transversal and recursive digital filters. It is frequently required that the frequency characteristic of a digital filter being synthesized should approximate that of the analog prototype with sufficient accuracy. The choice of a particular filter structure within the framework of the invariant impulse response method has a marked effect on the accuracy of approximation.

Let us compare the frequency characteristics of the two digital filters examined in Examples 15.6 and 15.7. Both correspond to an analog prototype whose frequency response is

$$K(j\omega) = 1/(1 + j\omega\tau) \tag{15.72}$$

On putting $\tau/\Delta = 5$ to make the matters more specific and on carrying out simple manipulations on the basis of Eqs. (15.72), (15.68) and (15.71), we can write the following expressions for the normalized amplitude response of the analog filter and of the two digital filters, recursive and transversal

$$\left|\frac{K(j\omega)}{K(j0)}\right|_a = \frac{1}{\sqrt{1 + 25\omega^2\Delta^2}} \tag{15.73}$$

$$\left|\frac{K(j\omega)}{K(j0)}\right|_r = \frac{0.1811}{\sqrt{1.6703 - 1.6375\cos\omega\Delta}} \tag{15.74}$$

$$\left|\frac{K(j\omega)}{K(j0)}\right|_t = \frac{\sqrt{2.120 + 2.7351\cos\omega\Delta + 1.3406\cos 2\omega\Delta}}{2.48903} \tag{15.75}$$

▲ Solve Problem 12

The values of $|K(j\omega)/K(j0)|$, as found from the above equations, are summarized in Table 15-1.

Table 15-1

$\omega\Delta$	Filter type		
	analog	digital recursive	digital transversal
0.0	1.0000	1.0000	1.0000
0.5	0.3714	0.3754	0.9200
1.0	0.1961	0.2046	0.7005
1.5	0.1322	0.1454	0.3963
2.0	0.0995	0.1182	0.1305
2.5	0.0797	0.1050	0.2234
3.0	0.0665	0.1000	0.3360

As is seen, both the recursive and the transversal digital filter has the frequency characteristics of a low-pass filter. However, the recursive filter is closer in properties to the analog prototype than the transversal filter.

Digital filter synthesis by digitization of the differential equation of the analog prototype. A digital filter structure approximately corresponding to a known analog network can be derived by digitizing the differential equation which describes the analog prototype. As an example of the procedure, we will go through the synthesis of a digital filter similar to a 2nd-order dynamic system

for which the relation between the output wave $y(t)$ and the input signal $x(t)$ is established by the differential equation

$$d^2y/dt^2 + 2\alpha\, dy/dt + \omega_0^2 y = x(t) \tag{15.76}$$

Suppose that the sampling interval is Δ, and consider two sets of samples, $\{y_k\}$ and $\{x_k\}$. If in Eq. (15.76) we replace the derivatives with their finite-difference expressions, the differential equation will turn into a difference equation

Higher-order finite-difference procedures may also be used

$$\frac{y_n - 2y_{n-1} + y_{n-2}}{\Delta^2} + 2\alpha\frac{y_n - y_{n-1}}{\Delta} + \omega_0^2 y_n = x_n \tag{15.77}$$

On re-grouping, we get

$$y_n = \frac{\Delta^2 x_n + 2(1 + \alpha\Delta)\, y_{n-1} - y_{n-2}}{A} \tag{15.78}$$

where $A = 1 - 2\alpha\Delta + \omega_0^2\Delta^2$.

The difference equation (15.78) defines the algorithm for the 2nd-order recursive filter which models an analog oscillatory system. This form of digital filter is called a *digital resonator.* If the coefficients have been properly chosen, a digital resonator can operate as a frequency-selective bandpass filter similar to a resonant (tuned) circuit.

● The digital resonator

The invariant frequency characteristic method. There are basic reasons why it is impossible to build a digital filter whose frequency behaviour would exactly replicate that of some analog network. One reason is that the frequency response of a digital filter is, as will be recalled, a periodic function of frequency with a period decided by the sampling interval Δ (Fig. 15.12).

As regards the similarity between (or the invariance of) the frequency behaviour of an analog and a digital filter, we may only

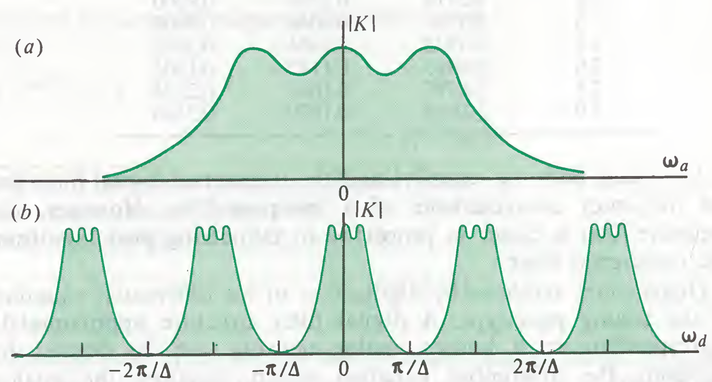


Fig. 15.12 Comparison of amplitude responses: (*a*) analog filter; (*b*) digital filter

require that all of the infinite frequency interval ω_a associated with the analog system be transformed into the frequency interval ω_d associated with the digital filter, such that

■ The principle of similarity between the frequency responses of analog and digital filters

$$-\pi/\Delta < \omega_d < \pi/\Delta \tag{15.79}$$

with the amplitude response retaining its general shape.

Let $K_a(p)$ be the transfer function of an analog filter, specified by a rational function in powers of the complex frequency p. Taking advantage of the fact that $z = \exp(p\Delta)$, we may write

$$p = (1/\Delta)\ln z \tag{15.80}$$

From this relation, however, we cannot deduce a physically realizable system function for the digital filter, because the substitution of (15.80) into the expression for $K_a(p)$ would result in a system function which is not the quotient of two polynomials. It is required to find a rational function of z which would possess the basic property of the transform (15.80), that is, which would map points on the unit circle in the z-plane into points on the $j\omega$-axis in the p-plane.

Among the various procedures, the following relation is widely used [35]

The function in (15.81) is called the bilinear transformation

$$p = (2/\Delta)(z - 1)/(z + 1) \tag{15.81}$$

which establishes a one-to-one correspondence between points on the unit circle in the z-plane and the entire imaginary axis in the p-plane. A salient feature of this form of mapping is as follows. By a change of variable, $z = \exp(j\omega_s\Delta)$, in (15.81), we get

$$j\omega_a = \frac{2}{\Delta}\,\frac{\exp(j\omega_s\Delta) - 1}{\exp(j\omega_s\Delta) + 1}$$

Hence, the following relation between the frequency variables ω_a and ω_s of the analog and the digital system respectively, emerges:

■ Relation between the frequency variables of an analog and a digital filter

$$\omega_a = \frac{2}{\Delta}\tan(\omega_s\Delta/2) \tag{15.82}$$

If the angular sampling rate is sufficiently high ($\omega_s\Delta \ll 1$), then, as can be readily seen from Eq. (15.82), $\omega_a \approx \omega_s$. Thus at low frequencies the characteristics of an analog filter and of a digital filter are about the same. In the general case, one should consider the scale transformation along the frequency axis of the digital filter described by Eq. (15.82).

Practically, the synthesis of a digital filter consists in that the

variable in the function $K_a(p)$ is changed in accord with Eq. (15.81). The system function thus obtained is a rational function, so the digital filtering algorithm may be written directly.

Example 15.8. *Synthesize a digital filter whose frequency behaviour is similar to that of a 2nd-order Butterworth analog low-pass filter. The cut-off frequency for the digital filter must be* $\omega_{c,d} = 1500\ \mathrm{s}^{-1}$. *The sampling rate is* $\omega_s = 10\,000\ \mathrm{s}^{-1}$.

To begin with, we find the sampling interval

$$\Delta = 2\pi/\omega_s = 6.2832 \times 10^{-4}\ \mathrm{s}$$

By Eq. (15.82), the cut-off frequency of the Butterworth analog filter similar to the digital filter being synthesized is

$$\omega_{c,a} = \frac{2}{\Delta}\tan\frac{\omega_{c,d}\Delta}{2} = 1621.9\ \mathrm{s}^{-1}$$

As will be recalled, the transfer function of a 2nd-order Butterworth analog filter, written in terms of the normalized complex frequency p_N, has the form (see Chap. 13)

$$K_a(p_N) = \frac{1}{p_N^2 + \sqrt{2}\,p_N + 1} \tag{15.83}$$

or, changing back to the true complex frequency,

$$K_a(p) = \frac{\omega_{c,a}^2}{p^2 + \sqrt{2}\,\omega_{c,a}\,p + \omega_{c,a}^2} \tag{15.84}$$

By a change of variable in (15.84) in accord with Eq. (15.81), the system function of the digital filter is

$$H(z) = \frac{\omega_{c,a}^2(z+1)^2}{\left[\left(\dfrac{2}{\Delta}\right)^2 + \sqrt{2}\,\omega_{c,a}\left(\dfrac{2}{\Delta}\right) + \omega_{c,a}^2\right]z^2 + 2\left[\omega_{c,a}^2 - \left(\dfrac{2}{\Delta}\right)^2\right] \atop + \left(\dfrac{2}{\Delta}\right)^2 - \sqrt{2}\left(\dfrac{2}{\Delta}\right)\omega_{c,a} + \omega_{c,a}^2} \tag{15.85}$$

On substituting the numerical values in the above expression, we get

$$H(z) = \frac{z^2 + 2z + 1}{7.6272z^2 - 5.7033z + 2.0761} \tag{15.86}$$

The effect of quantization on the performance of a digital filter. In the synthesis and design of digital filters it is important in some cases to consider the specific errors arising from signal quantization by which all quantities, both constant and time-varying, are turned into binary numbers of a finite length.

Quantization noise

Quantization results in a fairly wide range of effects. They are all examined in the literature on digital filtering [35]. Here we will be concerned with the *quantization* (or *quantizing*) *noise*, the simplest effect of all.

Let V_{max} be the maximum value of the analog signal applied to the input of the analog-to-digital converter of a digital filter, which does not yet cause an overflow in the arithmetic units of the filter. If m is the number of bits (word length) allocated to represent the input signal in the filter, it is then obvious that the input signal will be quantized into levels spaced apart by a quantization step of size

$$q_Q = V_{max}/2^m \tag{15.87}$$

The quantized levels represent the true instantaneous values of an analog signal accurate to what is known as the quantizing error which decreases as the step size, or the spacing between the quantizing levels, is reduced. So the quantized representation of the input signal, x_k, may be visualized as the sum of the true values $\tilde{x}_k$ and the errors n_k which constitute the quantization noise

Solve Problem 14

$$x_k = \tilde{x}_k + n_k \tag{15.88}$$

As theoretical and experimental studies have shown, in most cases of practical interest the error sequence $\{n_k\}$ is made up of statistically independent random variables each of which is uniformly distributed in the interval from $-q_Q/2$ to $q_Q/2$, and so the error sequence has zero mean and its variance is $\sigma_{in}^2 = q_Q^2/12$ (see Chap. 6).

The quantization noise present at the input of a digital filter is processed in the same manner as the valid signal. Let $\{n_{in,k}\}$ be the discrete sequence representing the input quantization noise. In order to find the lth sample in the output noise sequence $\{n_{out,k}\}$, we should evaluate the convolution of the input quantization noise with the impulse response of the filter

$$n_{out,l} = \sum_{i=0}^{\infty} h_i n_{in,l-i} \tag{15.89}$$

Hence the autocorrelation function of the output quantization noise is

$$K_{out}(m) = \sum_{i=0}^{\infty} n_{out,i} n_{out,i-m}$$

$$= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} h_j n_{in,i-j} h_{j-m} n_{in,i-j-m}$$

The sum is taken over the index i

$$= K_{in}(m)\sum_{j=0}^{\infty} h_j h_{j-m} \tag{15.90}$$

On setting $m=0$, the variance of the output noise is found to be

▲ Solve Problem 15

$$\sigma^2_{\text{out}} = K_{\text{out}}(0) = K_{\text{in}}(0) \sum_{j=0}^{\infty} h_j^2 = \frac{q_Q^2}{12} \sum_{j=0}^{\infty} h_j^2 \qquad (15.91)$$

As is seen, the output quantization noise increases as the roll-off of the impulse response of the filter slows down.

Summary

- In contrast to analog signals, discrete (sampled) signals are represented by sequences of samples at a discrete set of points.
- The spectrum of a discrete signal consists of an infinite number of repetitions of the spectrum of the original analog signal.
- The reconstruction of the original signal from a discrete sequence of samples by a real frequency filter inevitably introduces signal distortion.
- The number of amplitude harmonic coefficients that can be found with the aid of the discrete Fourier transform is half the number of samples.
- The z-transform permits discrete sequences to be investigated by the techniques of mathematical analysis applicable to continuous functions.
- The output sequence of a digital filter is the discrete convolution of the input signal and the impulse response of the filter.
- The frequency response of a digital filter is the Fourier transform of its impulse response and is a periodic function of frequency with a period equal to the sampling frequency (rate).
- The system function of a digital filter is the z-transform of its impulse response.
- According as the filtering algorithm is realized, digital filters are customarily classed into transversal or recursive.
- A recursive digital filter is stable if all the poles of its system function are located inside the unit circle with centre at $z=0$.
- In the synthesis of digital filters, it is convenient to use the impulse or frequency response of an analog prototype filter.
- The representation of quantized data as binary numbers of finite length leads to a specific source of error in the operation of digital filters, called the quantization noise.

Review Questions

1. Draw a block diagram for a pulse modulator. What is the difference between PAM and PWM signals?
2. What is the spectrum of the sampling sequence?
3. What is the cause of the distortion in the signal appearing at the output of a reconstructing low-pass filter, if its cut-off frequency is substantially higher than the frequency (rate) at which the input pulse sequence is sampled?
4. Over its periodicity interval a sampled signal is specified by 16 samples. What is the highest harmonic that can be found from these data with the aid of the discrete Fourier transform?

5. Formulate the conditions of convergence for the z-transform.
6. If the z-transform of a numeric sequence is known, how can the Fourier or Laplace transform of the corresponding discrete signal be found?
7. What is the number of bit positions (word length) ordinarily used in digital filters to represent sampled waveforms?
8. List the properties of the impulse response of a stationary linear digital filter. Define the condition that the impulse response of a physically realizable digital filter should satisfy.
9. Why is it that the representation of harmonic analog signals with a sequence of equidistant samples does not give a unique result?
10. Define the concept of the system function of a digital filter.
11. What are the differences between the system functions of transversal and recursive digital filters?
12. What is the basic difference between the impulse responses of transversal and recursive digital filters?
13. Name the advantages of the canonic form of digital filters.
14. What is the rationale of digital filter synthesis by the invariant impulse response method?
15. Why is it impossible to synthesize a digital filter whose frequency behaviour precisely replicates that of the analog prototype filter?
16. How does the shape of the impulse response of a digital filter affect the variance of the output quantization noise?

Problems

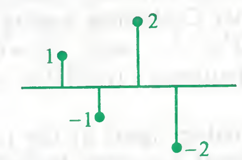
1. An analog signal $x(t)$ is a rectangular video pulse of duration $\tau_p = 2$ ms. Its sampled version is a discrete pulse sequence (a PAM signal) consisting of ten equidistant video pulses each of 5 μs duration. Find the spectrum of the pulse sequence.

2. Over its periodicity interval a discrete signal is specified by four equidistant samples: $\{x_k\} = (1,\ 0,\ -1,\ 0)$. Find the discrete Fourier transform coefficients of the signal.

3. Derive the equation describing the analog signal $x(t)$ reconstructed from the discrete Fourier transform coefficients of Problem 2.

4. A periodic discrete signal has the following harmonic amplitude coefficients: $C_0 = 0.5$, $C_1 = 1.5$ (the coefficients of the higher harmonics are zero). Find the sample values of the signal.

5. A sampled signal $\{x_k\}$ is specified by four samples. Find the z-transform of the signal:



6. Find the z-transform corresponding to an analog signal $x(t) = at\ (t > 0)$, where a is a constant.

7. Let the z-transform of a discrete signal $\{x_k\}$ have the form

$$X(z) = \frac{z^2 + 2z + 1}{z}$$

Find the samples of the discrete signal.

8. The impulse response of a digital filter

is specified by three nonzero samples: $\{h_k\} = (1, 0.5, 0.25)$. Find the system function and the frequency response of the digital filter.

9. Draw a block diagram of the digital filter realizing the following algorithm:

$$y_i = 1.75x_i - 0.55x_{i-1} + 0.25x_{i-2}$$

Find the system function.

10. A recursive digital filter realizes the following algorithm:

$$y_i = x_i + 0.5y_{i-1} - 0.75y_{i-2}$$

Analyse the digital filter for stability.

11. Calculate and plot the impulse responses of the digital filters described by the following difference equations:

(a) $y_i = 2.5x_i - 0.8y_{i-1}$

(b) $y_i = 2.5x_i + 0.8y_{i-1}$

12. Using the invariant impulse response method, synthesize a transversal digital filter similar to an integrating *RC* network. The impulse response of the digital filter is specified by four equidistant samples. Calculate the frequency response of the synthesized filter on setting $\Delta/\tau = 0.5$.

13. Using the invariant frequency response method, synthesize a digital filter whose frequency response is similar to that of a 2nd-order Chebyshev analog filter with $\omega_{c,a} = 2 \times 10^3\ s^{-1}$ and $\varepsilon = 0.8$. The sampling frequency (rate) is $\omega_s = 1.5 \times 10^4\ s^{-1}$.

14. The analog signal at the input of an analog-to-digital converter has a maximum amplitude $V_m = 50$ V. The samples are represented by eight-digit binary numbers. Calculate the variance of the quantization noise at the output of the analog-to-digital converter.

15. There is a recursive digital filter whose algorithm is $y_i = 0.45x_i + 0.95y_{i-1}$. The quantization noise at its input is quantized into levels spaced $q_Q = 0.5$ mV apart. Find the variance of the quantization noise at the output of the filter.

Advanced Problems

16. Think of representing a periodic discrete signal with the aid of a Fourier-Walsh transform in which the basis functions of the form $\exp(-j2\pi nk/N)$ used in the conventional discrete Fourier transform are replaced by Walsh functions.

17. Propose a digital filtering procedure based on the direct use of the discrete and inverse discrete Fourier transforms.

18. Using FORTRAN, write a fragment of the routine implementing the algorithm for recursive digital filtering.

19. Using the method of invariant impulse responses, synthesize a digital filter corresponding to an analog lossy resonant circuit whose impulse response has the form

$$h(t) = \exp(-\alpha t)\cos\omega_0 t$$

20. In mathematics, a sequence of numbers related such that $x_i = x_{i-1} + x_{i-2}$ is called the Fibonacci sequence. Find the first ten Fibonacci numbers on setting $x_0 = 0$ and $x_1 = 1$. Propose an analytic procedure for finding Fibonacci numbers of any order. Connect this problem to that of the stability of recursive digital filters.

Chapter 16
Optimum Linear Signal Filtering

Noise and interference control and suppression are an important task in many branches of telecommunications. The high noise immunity of telecommunication systems can in principle be assured in two ways. One consists in improving the structure of the signals being transmitted. An example is offered by the Barker sequences examined in Chap. 3. The other consists in developing equipment capable of signal detection in noise.

This chapter will be concerned with signal detection in noise by stationary linear systems acting as frequency filters. A frequency-selective system which processes the sum of a signal and noise in the best possible manner is called an *optimum filter*.

The optimum filter

The meaning that is implied in the concept of optimality varies among other things, according as the signal waveform is known in advance or not. If the signal waveform is known, an optimal filter is expected to detect the signal with a maximally attainable probability. If, on the other hand, the signal waveform is not known, an optimal filter must separate the signal from its mixture with noise as best as it can. In the latter case, an optimum filter is customarily taken as that which minimizes the root-mean-square error between the actual signal and the filter output.

The difference between the detection of a known and an unknown signal waveform

16.1 Optimum Linear Filtering of Known Signals

The pages that follow will set forth the theory used to build linear optimum filters in both cases.

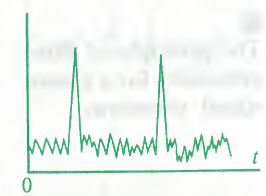
The task of the optimum detection of a known signal in noise arises, for example, in radar. Here, the received signal $s_r(t)$ is an exact, scaled-down replica of the transmitted signal $s_t(t)$, that is

$$s_r(t) = As_t(t - \tau) \tag{16.1}$$

where $A \ll 1$.

The received signal may be very small and comparable in amplitude with the effective noise voltage present at the receiver input.

A radar receiver does two jobs: (a) it detects the signal, that is, it establishes the fact that the reflected signal (the target echo) is present in the received wave; and (b) it determines the delay time τ, which is a measure of the distance to the target.



Signal "spikes" above the noise level

A specific aspect in the operation of a radar is that in processing the received signal the signal waveform need not be retained. In fact, it is desired that in the course of processing the signal be transformed in such a way that its application to the filter input would produce at some instant an appreciable "spike" in the

instantaneous value of the output waveform. The noise present in the received waveform is usually Gaussian, so the probability that it, too, will give rise to noticeable "spikes" is negligible. Therefore, the appearance in the output waveform of a spike substantially exceeding the effective noise voltage indicates the presence of the signal at the receiver input with a high probability.

■ **The assumption that noise is Gaussian**

The signal-to-noise ratio at the output of a linear filter. We assume that the sum of the signal and noise is processed by a system which is a stationary linear filter whose frequency response is

$$K(j\omega) = |K(j\omega)| \exp[j\varphi_K(\omega)]$$

The spectrum of a known signal $s_{in}(t)$ at the filter input is

$$S_{in}(\omega) = |S_{in}(\omega)| \exp[j\varphi_S(\omega)]$$

Using spectral analysis, we can find the signal at the filter output at any time t_0

$$s_{out}(t_0) = \frac{1}{2\pi}\int_{-\infty}^{\infty} |S_{in}||K| \exp[j(\omega t_0 + \varphi_S + \varphi_K)]\,d\omega \qquad (16.2)$$

Suppose that the noise accompanying the signal at the filter input is white noise for which the power spectrum W_0 is the same at all frequencies. Then the variance of the noise at the filter output will be

$$\sigma^2_{out} = \frac{W_0}{2\pi}\int_{-\infty}^{\infty} |K|^2\,d\omega \qquad (16.3)$$

● **The signal-to-noise ratio (SNR)**

The ratio of the magnitude of the instantaneous value of the signal at time t_0 to the root-mean-square value of noise is known as the *signal-to-noise ratio* (SNR) at the filter output. Designating it as q and using Eqs. (16.2) and we get

$$q = \frac{s_{out}(t_0)}{\sigma_{out}} = \frac{\left|\dfrac{1}{2\pi}\displaystyle\int_{-\infty}^{\infty} |S_{in}||K| e^{j(\omega t_0 + \varphi_S + \varphi_K)}\,d\omega\right|}{\left(\dfrac{W_0}{2\pi}\displaystyle\int_{-\infty}^{\infty} |K|^2\,d\omega\right)^{1/2}} \qquad (16.4)$$

■ **The principle of filter optimality for a known signal waveform**

An optimum filter will be one with $K_{opt}(j\omega)$ which maximizes the signal-to-noise ratio at some time t_0. This form of optimum filter is also called the *matched filter*, because it is matched to a known signal.

● **The matched filter**

The frequency response of a matched filter. The task of finding $K_{opt}(j\omega)$ is handled on the basis of the Cauchy-Buniakovsky

inequality. By this inequality

The equation is generalized to functions which take on complex values

$$\left|\int F_1(x) F_2^*(x)\,dx\right| \leqslant \left(\int |F_1|^2\,dx \int |F_2|^2\,dx\right)^{1/2} \qquad (16.5)$$

for arbitrary functions $F_1(x)$ and $F_2(x)$. In Eq. (16.5) the "equals" sign applies only when $F_1(x) = CF_2(x)$, where C is a constant.

From comparison of the left-hand side of (16.5) with the numerator of (16.4), let us put

$$F_1(\omega) = |S_{in}| \exp(j\varphi_S) \qquad (16.6)$$

$$F_2^*(\omega) = |K| \exp[j(\omega t_0 + \varphi_K)]$$

The numerator in (16.4) will be a maximum if F_1 and F_2 are proportional to each other, that is,

$$|S_{in}| \exp(j\varphi_S) = C|K| \exp[-j(\omega t_0 + \varphi_K)]$$

The equality of two complex numbers implies that their magnitudes and arguments are respectively equal. In our case,

$$|S_{in}| = C|K| \qquad (16.7)$$

$$\varphi_S = -\omega t_0 - \varphi_K$$

Hence, we immediately deduce the expression for the frequency response of a matched (optimum) filter

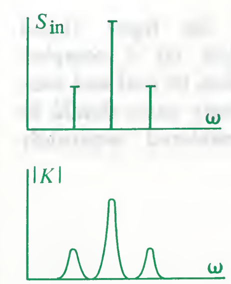
$$K_{opt}(j\omega) = B|S_{in}| \exp(-j\varphi_S) \exp(-j\omega t_0) \qquad (16.8)$$

where $B = 1/C$ is an arbitrary coefficient of proportionality. It is convenient to re-write Eq. (16.8) as

$$K_{opt}(j\omega) = BS^*_{in}(\omega) \exp(-j\omega t_0) \qquad (16.9)$$

The result holds if the noise at the output is white

Thus, the frequency response of a matched filter is completely defined by the spectrum of the signal which the filter is intended to detect. The proportionality factor B in (16.9) determines the general level of gain due to the filter. The value of t_0 only enters the expression for the phase response of the filter. From the general properties of the spectral representation of signals we know that the exponential factor $\exp(-j\omega t_0)$ is an indication that the filter response is shifted along the time axis by t_0.



The signal spectrum and the amplitude response of a comb filter

Physical interpretation of the frequency response of the optimum filter. A filter intended to extract a known signal from a white noise mixture will pass the harmonic waves at the frequencies where the spectrum of the signal is nonzero. Also, naturally, the amplitude response of the filter will be proportional to the magnitude of the spectrum of the signal, that is, to the contribution that a particular frequency interval makes to the output signal. If the input signal

has a discrete spectrum (such as when the signal is periodic), the above principle leads to comb filters widely used in telecommunications.

The matched filter whose behaviour is defined by Eq. (16.9) functions similarly to a comb filter. However, it assures a still higher probability of signal detection because the phase characteristic of the signal spectrum is utilized to a better advantage. To demonstrate, the signal at the output of a matched filter

$$s_{out}(t) = \frac{B}{2\pi} \int_{-\infty}^{\infty} S_{in} S^*_{in} \exp[j\omega(t - t_0)]\, d\omega \tag{16.10}$$

is obviously, at an absolute maximum

Rayleigh's theorem is used

$$s_{out,max} = \frac{B}{2\pi} \int_{-\infty}^{\infty} |S_{in}|^2 d\omega = BE_s \tag{16.11}$$

(where E_s is the signal energy) at time t_0 when all elementary spectral components of the input wave are combined at the output coherently, that is, in the same time phase.

In optimal filtering the spectral components are combined coherently

Thus, *matched filtering calls for the maintenance of proper phase relations between the individual spectral components of the detected signal.*

The impulse response of a matched filter. If we know the frequency response of a matched filter, we can find its impulse response by taking the inverse Fourier transform of Eq. (16.9):

$$h(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} K_{opt}(j\omega) \exp(j\omega t)\, d\omega$$

$$= \frac{B}{2\pi} \int_{-\infty}^{\infty} S^*_{in}(\omega) \exp[j\omega(t - t_0)]\, d\omega \tag{16.12}$$

If the input signal takes on a complex value, its real and imaginary parts should be considered separately

As will be recalled, any real signal has the property that $S^*_{in}(\omega) = S_{in}(-\omega)$. Therefore,

$$h(t) = \frac{B}{2\pi} \int_{-\infty}^{\infty} S_{in}(-\omega) \exp[j\omega(t - t_0)]\, d\omega$$

$$= -\frac{B}{2\pi} \int_{\infty}^{-\infty} S_{in}(\omega') \exp[j\omega'(t_0 - t)]\, d\omega'$$

$$= Bs_{in}(t_0 - t) \tag{16.13}$$

Thus, we have proved that the impulse response of a matched filter is a scaled replica of the input signal, but this replica is a mirror image of the input signal on the time axis–this is

indicated by the "–" sign of t in Eq. (16.13). Also, the impulse response of the filter is shifted to the right, so it is delayed from the signal $s_{in}(-t)$ by t_0.

The construction of the impulse response of the matched filter for a specific pulse signal $s_{in}(t)$ of finite duration is illustrated in Fig. 16.1.

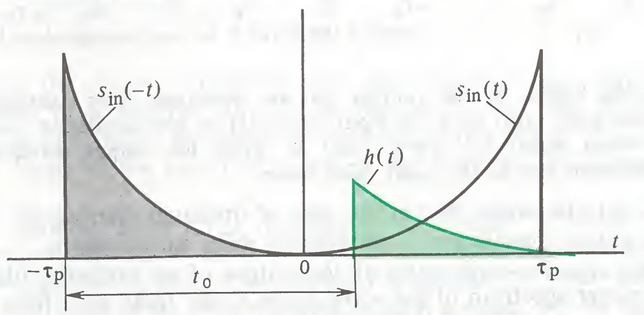


Fig. 16.1 Construction of the impulse response for a matched filter

Solve Problem 1

Referring to Fig. 16.1, we can formulate a necessary, but not a sufficient condition for the physical realizability of a matched filter: The time parameter t_0 defining the instant when the output signal has a maximal instantaneous value must be *not less than the duration of detected signal.* Otherwise, the impulse response of the system will be nonzero at $t < 0$, that is, before the delta-impulse is applied to the filter input.

This type of matched filter can only be synthesized for signals of finite energy, such as pulses

The implication of the condition is this: If the output signal of an optimum filter is to have a maximally possible instantaneous value, the filter must utilize the energy of the *entire* input signal.

The waveform of the output signal of a matched filter. Suppose that a filter receives a noise-free signal $s_{in}(t)$ to which it is matched. Let us analyse the waveform of the output signal. To do this, we write on the basis of (16.10)

$$s_{out}(t) = \frac{B}{2\pi} \int_{-\infty}^{\infty} |S_{in}(\omega)|^2 \exp[j\omega(t - t_0)]\, d\omega \tag{16.14}$$

Going back to Eq. (3.24), we can see that the output signal coincides to within the proportionality factor B with the autocorrelation function of the input signal, shifted backwards by t_0, that is,

$$s_{out}(t) = BK_s(t - t_0) \tag{16.15}$$

As an example, Fig. 16.2 shows a plot of the signal at the output of some specific matched filter.

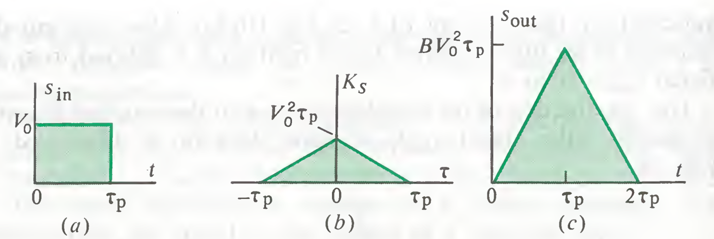


Fig. 16.2 Signal at the output of an optimum filter matched to a rectangular video pulse: (*a*) input signal; (*b*) its autocorrelation function; (*c*) output signal for $t_0 = \tau_p$, that is, when the output waveform is a maximum just as the input pulse ceases

Waveform distortion does not interfere with the detection of a signal in noise

It is to be noted that in the case of optimum filtering the input and output signals may substantially differ in waveform.

The signal-to-noise ratio at the output of an optimum filter. If the power spectrum of the white noise at the input of a filter with the frequency response of Eq. (16.9) is W_0, then the power spectrum of the output noise will be

$$W_{out}(\omega) = W_0 |K_{opt}(j\omega)|^2 = W_0 B^2 |S_{in}(\omega)|^2 \tag{16.16}$$

The variance of the output noise is found by integrating Eq. (16.16) over all frequencies. It is connected to the energy of the input signal, E_s, by a relation of the form

$$\sigma^2_{out} = \frac{W_0 B^2}{2\pi} \int_{-\infty}^{\infty} |S_{in}(\omega)|^2 \, d\omega = W_0 B^2 E_s$$

Hence, using the expression for the maximum response produced by the filter (see Eq. (16.11)), we find the maximum attainable signal-to-noise ratio

$$q_{max} = |s_{out,\,max}| / \sigma_{out} = \sqrt{E_s/W_0} \tag{16.17}$$

The maximum attainable SNR

It is important to stress that Eq. (16.17) establishes the fundamental limit of signal detectability by devices intended to detect known signals mixed with white noise of specified intensity.

Example 16.1. *The signal being detected is a rectangular radio pulse of amplitude V_0 and duration $\tau_p = 10$ μs. The white noise at the filter input has the power spectral density $W_0 = 3 \times 10^{-18}$ V²s. Find the minimum value of V_0 at which the signal can still be detected, if the receiver reliably indicates the presence of the signal with a signal-to-noise ratio of $q = 3$.*

On the basis of (16.17), the required signal-to-noise ratio will be realized if the signal energy is $E_s = 9W_0$. Since for a rectangular radio pulse, $E_s = V_0^2\tau_p/2$, then $V_{0,\,min} = \sqrt{18W_0/\tau_p} = 2.32 \times 10^{-6}$ V.

A remarkable property of matched filtering is that the probability of detecting a signal depends above all on the energy in the signal. Notably, we can always provide for the reliable detection of a signal with a very small amplitude, if we extend the duration of the pulse in a suitable manner. This will of course slow down the transmission of information over the channel.

Work Problem 2

16.2 Implementation of Matched Filters

From the above expressions for the frequency and impulse responses of a matched filter, it is possible to find the physical structure of a device intended for the optimal filtering of a known signal. Some of the procedures used in the synthesis are discussed below.

The matched filter for a rectangular video pulse. Consider an elementary pulse signal $s_{in}(t)$ which is a rectangular video pulse of known duration τ_p and of an arbitrary amplitude V_0. Our search for the structure of the filter matched to such a signal will be based on the spectral principle. To begin with, we find the spectrum of the signal at the filter input:

$$S_{in}(\omega) = \int_{-\infty}^{\infty} s_{in}(t) \exp(-j\omega t)\, dt = V_0 \int_0^{\tau_p} \exp(-j\omega t)\, dt$$
$$= (V_0/j\omega)[1 - \exp(-j\omega\tau_p)] \tag{16.18}$$

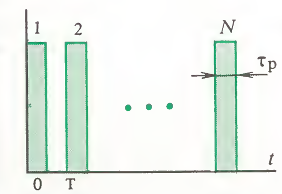
Here the frequency response is not a rational function, so the corresponding filter cannot be a lumped-parameter network

Hence, basing ourselves on Eq. (16.9), we find the frequency response of the matched filter by setting $t_0 = \tau_p$, which means that the output of the filter is a maximum at the instant when the pulse ceases:

$$K_{opt}(j\omega) = B \frac{1 - \exp(j\omega\tau_p)}{-j\omega} \exp(-j\omega\tau_p)$$
$$= B(1/j\omega)[1 - \exp(-j\omega\tau_p)] \tag{16.19}$$

The result thus obtained is sufficient for a matched filter to be synthesized. As follows from Eq. (16.19), this filter must be a cascaded connection of three linear elements: (a) a scaling amplifier with gain B, (b) an ideal integrator, and (c) a network with a frequency response, $K(j\omega) = 1 - \exp(-j\omega\tau_p)$ which can be implemented with a delay unit to delay the signal for time τ_p, an inverter to change the sign of the signal, and an adder. In block-diagram form, the filter thus synthesized is shown in Fig. 16.3.

The matched filter for a train of identical video pulses. In radar the waveforms coming from the amplitude detector of the receiver are often combined into trains so as to build up the power in the



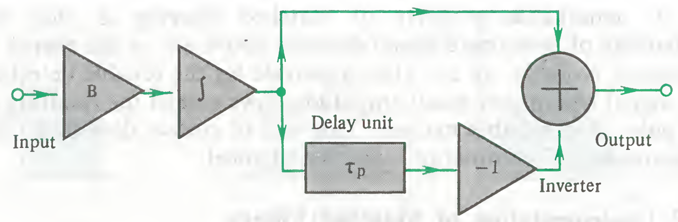


Fig. 16.3 Matched filter for a rectangular video pulse

signal. Suppose that such a train consists of N identical video pulses of duration τ_p each; the spacing between the pulses is T. If $S_0(\omega)$ is the spectrum of one pulse, then the spectrum of the train will be

$$S_t(\omega) = S_0(1 + e^{-j\omega T} + e^{-j2\omega T} + \ldots + e^{-j(N-1)\omega T}) \tag{16.20}$$

In synthesizing a matched filter for a pulse train, we require that the response be a maximum just as the last pulse in the train ceases. Hence,

$$t_0 = (N-1)T + \tau_p$$

By applying Eq. (16.9), we obtain the following expression for the frequency response of the optimum filter:

$$K_{opt}(j\omega) = BS_0^* e^{-j\omega\tau_p}(1 + e^{j\omega T} + e^{j2\omega T} + \ldots + e^{j(N-1)\omega T})$$
$$= K_{0,opt}(j\omega)(1 + e^{-j\omega T} + e^{j2\omega T} + \ldots + e^{-j(N-1)\omega T}) \tag{16.21}$$

where $K_{0,opt}(j\omega)$ is the frequency response of the optimum filter for a single video pulse.

The structure of the sought matched filter results directly from Eq. (16.21). In block-diagram form, it appears in Fig. 16.4. As is seen, a matched filter for a single video pulse is located at the input.

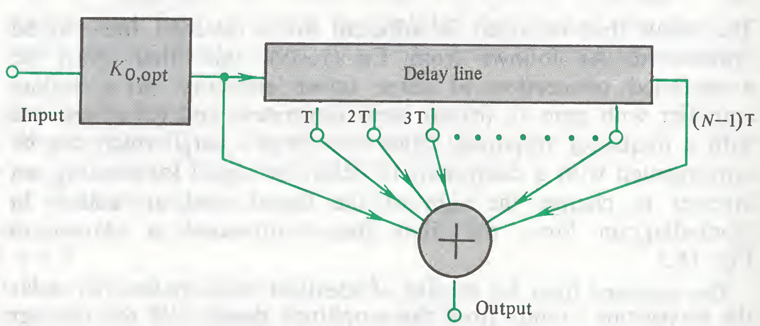


Fig. 16.4 Matched filter for a train of video pulses

The basis of the entire device is a multi-tap delay line which delays the signals for T, $2T$, ..., and $(N-1)T$. From all the taps, the signals go to the adder. It is an easy matter to see that the response at the adder output will be a maximum when the signals from all the pulses in the train are present at its inputs simultaneously.

The use of complex signals in the form of trains will naturally slow down the rate of data output as compared with single pulses

Thus, the above matched filter "compresses" the train by deriving a single response of a maximum amplitude from the entire sequence of pulses making up the train. The performance of the device improves with increasing length of the train.

Practical radar detectors also include a nonlinear threshold element whose input is connected to the output of the adder in the matched filter. The threshold level somewhat exceeds the root-mean-square level of noise in the no-signal condition. When the maximum spike of the filter output signal reaches the threshold level, a control signal is routed to the display unit, thereby indicating the presence of the pulses reflected from a target.

The matched filter for a rectangular radio pulse. Let the signal being detected be a rectangular radio pulse such that

$$s_{in}(t) = \begin{cases} 0, & t < 0 \\ V_0 \sin\omega_0 t, & 0 \leqslant t \leqslant \tau_p \\ 0, & t > \tau_p \end{cases} \tag{16.22}$$

We set out to synthesize a filter matched to such a signal by using our knowledge about the impulse response of the filter. As has been shown earlier, the impulse response of a matched filter is

$$h(t) = Bs_{in}(t_0 - t)$$

Let us put, as before, $t_0 = \tau_p$, and deem for simplicity that the pulse duration is a multiple of the period of the r.f. carrier and so $\sin\omega_0\tau_p = 0$, and $\cos\omega_0\tau_p = 1$. Then,

$$h(t) = \begin{cases} 0, & t < 0 \\ B\sin\omega_0 t, & 0 \leqslant t \leqslant \tau_p \\ 0, & t > \tau_p \end{cases} \tag{16.23}$$

If the oscillatory element has a high Q, the exponential decay of the amplitude with time may be neglected

As is seen, the impulse response of a matched filter replicates the input signal to within the amplitude factor. This impulse response can be approximately realized with the structure shown in block-diagram form in Fig. 16.5.

As is seen, at the filter input there is an oscillatory element, such as a high-Q resonant circuit whose impulse response is

$$h(t) = \begin{cases} 0, & t < 0 \\ H\sin\omega_0 t, & t > 0 \end{cases}$$

where H is a constant.

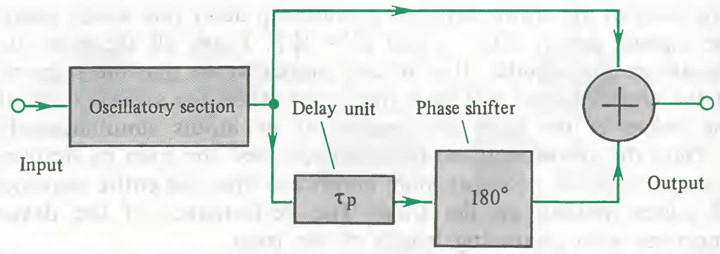


Fig. 16.5 Matched filter for a rectangular radio pulse

For the impulse response of the matched filter to be zero at $t > \tau_p$, the filter includes an adder one input of which receives the signal from the oscillatory element directly, and the other does so via a delay unit which delays the signal for τ_p. Connected in series with the delay unit is a phase shifter which reverses the phase of the signal. With this arrangement, the adder inputs receive, beginning at $t = \tau_p$, two harmonic waves which have the same amplitude, but are opposite in phase. The result is a zero signal at the adder output.

The matched filter for Barker sequences. A matched filter can be far more effective in signal compression, if the input consists of Barker sequences instead of simple trains of pulses. As is pointed out in Chap. 3, an advantage of Barker sequences is that the main lobe of the autocorrelation function has a very high value, whereas the side lobes are minimal.

Barker sequences may be of length $M = 2$, 3, 4, 5, 7, 11, 13 bits

A block diagram of a matched filter for M-digit Barker sequences is shown in Fig. 16.6. The filter utilizes the principle of phase-shift-keying in which, as will be recalled, the input signal consists of a sequence of chopped harmonic waves whose phase angles φ_1, φ_2, ..., φ_M are either zero or 180° (see Fig. 3.7).

In synthesizing this form of matched filter, we proceed from the

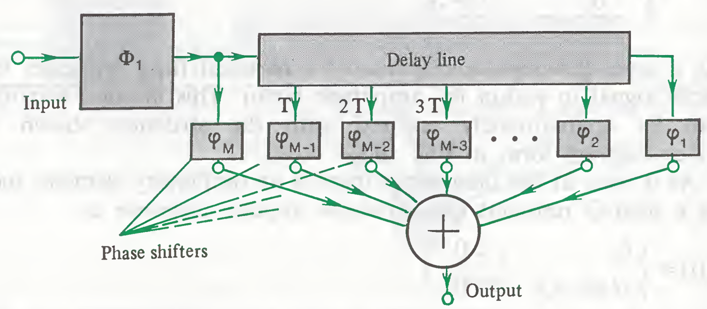


Fig. 16.6 Matched filter for a Barker sequence

fact that its impulse response must be a "mirror image" of the detected signal, with its elements following in a reverse time order.

At the input of the matched filter there is an auxiliary filter F_1 matched to one element of the complex PSK signal, that is, to a rectangular radio pulse. Under the action of the input delta-impulse, the filter output delivers a radio pulse with a rectangular envelope. This pulse is applied to a tapped delay line which usually is a distributed-parameter (wave) structure. The delay time between the taps is equal to the duration T of each element of the complex signal.

▲ Solve Problem 3

For the matched filter to operate properly, it is essential that the sequence of phase angles φ_M, φ_{M-1}, ..., φ_1 (Fig. 16.6) should correspond to the phase angles of the individual pulses of the Barker sequence as counted backwards.

Travelling along the delay line, the rectangular radio pulse consecutively excites the inputs of the adder, and this delivers a "mirror image" of the signal being detected.

The matched filter for LFM-pulses. Equation (16.17) tells us that the limiting value of the signal-to-noise ratio at the output of a matched filter depends on the energy in the signal, whatever its waveform may be. In practice, however, the task is usually not only to detect the presence of a signal, but also to measure some parameters of the signal, say its time position. In such cases, preference is given to signals whose autocorrelation function has a sharp maximum (see Chap. 3).

Among the several signals possessing this property, those especially widely used are linearly frequency-modulated (LFM) pulses. Their theory is set forth in Chap. 4. It has been shown that if an LFM pulse of the form

$$s_{in}(t) = \begin{cases} 0, & t < -\tau_p/2 \\ V_0 \cos(\omega_0 t + \mu t^2/2), & -\tau_p/2 < t < \tau_p/2 \\ 0, & t > \tau_p/2 \end{cases}$$

has a large bandwidth-duration product, $\mu\tau_p^2 \gg 1$, then its spectrum $S_{in}(\omega) = |S_{in}| \exp[j\phi(\omega)]$ within the bandwidth $\Delta\omega = \mu\tau_p$ has a practically constant magnitude

$$|S_{in}| = V_0 \sqrt{\pi/2\mu}$$

and the argument is a quadratic function of frequency:

$$\Phi(\omega) = -(\omega - \omega_0)^2/2\mu \tag{16.24}$$

Hence comes the requirements that should be satisfied by the frequency characteristic of the filter matched to an LFM-signal: For the output response to be a maximum at some time t_0, it is essential that the amplitude response of the filter should be constant in the frequency interval $(\omega_0 - \Delta\omega/2, \omega_0 + \Delta\omega/2)$ and that

its phase response should be

$$\varphi_K(\omega) = -\omega t_0 + (\omega - \omega_0)^2/2\mu \tag{16.25}$$

The phase response of the matched filter for an LFM signal

The first term on the right-hand side of Eq. (16.25) is responsible for the delay of the output signal as a whole for a time t_0, the second (quadratic) term compensates for the phase shifts between the individual components of the signal and thus assures that they are combined coherently at the output.

The quadratic character of the phase response of the matched filter for an LFM signal can be deduced from the following qualitative considerations. In the course of the intrapulse modulation the instantaneous frequency of the signal varies linearly as

$$\omega(t) = \omega_0 + \mu t$$

in the time interval $(-\tau_p/2, \tau_p/2)$. To each time t within the pulse duration there corresponds its own narrowband (quasiharmonic) signal which is delayed by the filter for the time interval equal to the group (envelope) delay time (see Chap. 8):

$$T_g = -d\varphi_K/d\omega = t_0 - (\omega - \omega_0)/\mu = t_0 - t \tag{16.26}$$

In order to find the instant when the individual spectral components appear at the output, we should add to the above time the amount t, that is, the time when the spectral components appear at the input. Hence we conclude that all the spectral components of the LFM signal appear at the filter output at the same time t_0.

The signal appearing at the output of the matched filter replicates to within an arbitrary amplitude factor A the autocorrelation function of the LFM pulse (see Eqs. (4.54) and (16.15)):

$$s_{out} = A\frac{\sin\dfrac{\mu\tau_p}{2}(t-t_0)}{\dfrac{\mu\tau_p}{2}(t-t_0)}\cos\omega_0(t-t_0) \tag{16.27}$$

A plot of the above signal has been given in Fig. 4.10. It is easy to see that the width of the main lobe of such a signal, as measured between zero points, is equal to

Solve Problem 4

$$\tau_{out} = 4\pi/\Delta\omega = 4\pi/\mu\tau_p$$

Therefore the compression ratio of an LFM pulse provided by the

matched filter

$$K_{comp} = \tau_p/\tau_{out} = \mu\tau_p^2/4\pi = \frac{\text{bandwidth} \times \text{duration}}{4\pi} \tag{16.28}$$

The compression ratio for a pulse

is directly proportional to the bandwidth-duration product of the LFM signal.

In constructing actual matched filters, advantage is taken of a physical phenomenon consisting in that elastic ultrasonic waves propagating in solids display what is known as dispersion, that is, variations in the velocity of propagation with frequency. Through the proper choice of the form of dispersion, it is possible to secure that the phase response has the form defined in Eq. (16.24). In sketch form, the arrangement of such a filter is shown in Fig. 16.7.

Delay lines are often made of aluminium alloys

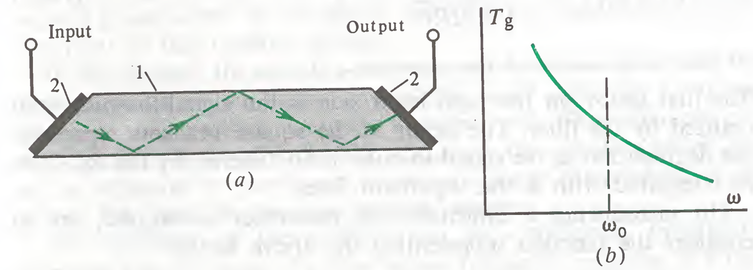
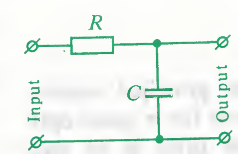


Fig. 16.7 Distributed-parameter filter matched to an LEM signal: (*a*) sketch; (*b*) frequency dependence of group delay in the acoustic line; (*1*) acoustic line; (*2*) electromechanical transducers

In contrast to the optimum filtering of trains of video pulses, the matched filtering of LFM pulses is usually effected at the carrier frequency or the intermediate frequency of the receiver, that is, ahead of the amplitude detector. This avoids the undesirable suppression of a weak signal by a strong interference which may arise when the sum of the signal and noise is subjected to nonlinear transformation.

An advantage of signal processing ahead of the detector

Quasi-optimum filters. In some cases, a known signal mixed with a Gaussian noise can be detected with sufficient effectiveness, using filters far simpler in design than optimum filters. They are customarily called *quasi-optimum filters*.



The theory of quasi-optimal signal filtering has been developed by V. I. Siforov (USSR)

Consider an integrating *RC* two-port whose input simultaneously receives white noise of power spectrum W_0 and a rectangular video pulse of amplitude V_0 and duration τ_p. Since the two-port is a linear network, the passage of the signal may be considered separately from that of the noise.

The output signal is a maximum at the instant when the pulse

ceases, and is equal to

$$s_{\text{out, max}} = V_0[1 - \exp(-\tau_p/RC)]$$

On the other hand, the variance of the output noise of the RC-network driven by white noise is defined (see Chap. 10) by

$$\sigma_{\text{out}}^2 = W_0/2RC$$

Hence, the maximum signal-to-noise ratio at the output of the RC-network is

$$q_{RC} = \frac{V_0[1 - \exp(-\tau_p RC)]}{\sqrt{W_0/2RC}} \tag{16.29}$$

Noting that the energy in the video pulse in question is $E_s = V_0^2\tau_p$, we may re-write (16.29) as follows:

$$q_{RC} = \sqrt{E_s/W_0}\left[\frac{1 - \exp(-\tau_p/RC)}{\sqrt{\tau_p/2RC}}\right] \tag{16.30}$$

The first factor on the right-hand side is the signal-to-noise ratio realized by the filter. The factor in the square brackets represents the degradation in the signal-to-noise ratio suffered by the RC-filter as compared with a true optimum filter.

On introducing a dimensionless parameter $x = \tau_p/RC$, let us consider the function representing the above factor:

$$k(x) = \frac{1 - \exp(-x)}{\sqrt{x/2}} \tag{16.31}$$

Its plot appears in Fig. 16.8. As is seen, at $x = 1.25$, $k(x)$ is at a maximum equal to 0.90.

The principal requirement for a quasi-optimal filter is to pass without attenuation the frequencies where the bulk of signal energy is concentrated

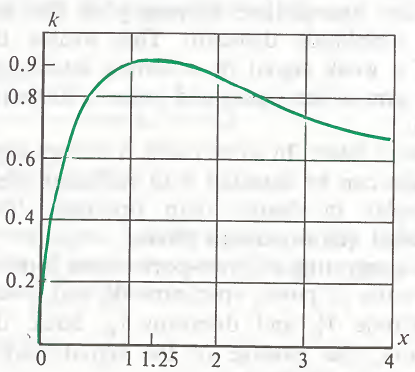


Fig. 16.8 Impairment in the signal-to-noise ratio of an RC filter in comparison with a matched filter

Thus, through the proper choice of the time constant for the RC-network, it is possible to build a very simple quasi-optimum filter for which the signal-to-noise ratio is only 10% lower than it is for a matched filter.

It should be noted that quasi-optimum filters of acceptable performance can be built only for relatively simple signals with a small bandwidth-duration product.

16.3 Optimum Filtering of Random Signals

In practice it is usual that the exact form of the signal is not known in advance. Examples are the real signals sent into the communication channel from sources such as a microphone, a TV camera, etc. In such cases, the instantaneous values of the signal should be treated as typical realizations of a stationary, ergodic ensemble, and the sole information about the entire set of likely signals is contained in the power spectrum (or the autocorrelation function) of this random process.

In the channel, the signal is accompanied by noise. As a rule, the power spectra of the signal and of the noise differ to a varying degree in their position along the frequency axis. This difference can be utilized in order to formulate and solve the problem which has as its objective to develop a stationary linear filter which could detect the signal in the best possible manner.

The statement of the problem and the criterion of optimality. Suppose that the input of a filter whose frequency response is $K(j\omega)$ is fed simultaneously two random signals whose realizations are $x(t)$ and $z(t)$. Let $x(t)$ be the wanted signal, and $z(t)$ be the noise. Both are realizations of two stationary random processes, $X(t)$ and $Z(t)$, respectively. It is presumed that these two random processes are uncorrelated and are defined by their power spectra, $W_x(\omega)$ and $W_z(\omega)$.

The realization $y(t)$ of the output signal of the filter is not an exact replica of the input signal $x(t)$, but differs from it by an amount known as the *random error*

$$e(t) = x(t) - y(t) \tag{16.32}$$

The optimum filter is then taken as that for which the error variance is a minimum.

This type of filter is said to minimize the mean squared error of signal reproduction

Relation of the error variance to the power spectrum. If $W_e(\omega)$ is the power spectrum of the error, then its variance is

$$\sigma_e^2 = \frac{1}{2\pi}\int_{-\infty}^{\infty} W_e(\omega)\,d\omega \tag{16.33}$$

Let us relate $W_e(\omega)$ to $W_x(\omega)$ and $W_z(\omega)$. For this purpose, we will

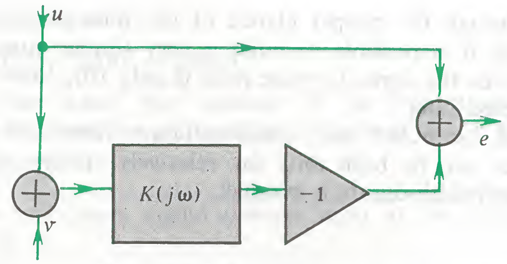


Fig. 16.9 Explaining the derivation of the error signal

consider the block diagram of an imaginary device whose output is the realization of the error, $e(t)$, (Fig. 16.9).

Since, by the statement of the problem, the processes $X(t)$ and $Z(t)$ are uncorrelated, the powers of the random signals reaching the output over each of the two available channels are combined, and so

$$W_e(\omega) = |K(j\omega)|^2 W_z(\omega) + |1 - K(j\omega)|^2 W_x(\omega) \tag{16.34}$$

Let us write the frequency response of the filter in exponential form:

$$K(j\omega) = H_k(\omega) \exp j\varphi_K(\omega)$$

and investigate the term $|1 - K(j\omega)|^2$ on the right-hand side of Eq. (16.34). Obviously,

$$|1 - K(j\omega)|^2 = H_k^2 - 2H_k \cos\varphi_K + 1$$

this quantity being a minimum when $\varphi_K = 0$. Thus, the optimum filter in question should introduce a zero phase shift at all frequencies.

The condition imposed on the phase response of an optimum filter

Noting this, we obtain the following expression for the error variance:

$$\sigma_e^2 = \frac{1}{2\pi} \int_{-\infty}^{\infty} [(H_k - 1)^2 W_x + H_k^2 W_z]\, d\omega \tag{16.35}$$

Minimization of the error variance. By carrying out simple identity transformations, it is convenient to write Eq. (16.35) as

The theory of optimum filtering as applied to random signals was developed by A. N. Kolmogorov and N. Wiener in the 1940s

$$\sigma_e^2 = \frac{1}{2\pi} \int_{-\infty}^{\infty} \left[\left(\sqrt{W_x + W_z}\, H_k - \frac{W_x}{\sqrt{W_x + W_z}}\right)^2 + \frac{W_x W_z}{W_x + W_z}\right] d\omega \tag{16.36}$$

The magnitude of the frequency response, $H(\omega)$, appears only in the integrand term within the parentheses. This term is nonnegative,

therefore the error variance is a minimum when

$$\sqrt{W_x + W_z}\, H_k - \frac{W_x(\omega)}{\sqrt{W_x + W_z}} = 0$$

Hence,

$$H_{k,\,\mathrm{opt}}(\omega) = \frac{W_x}{W_x(\omega) + W_z(\omega)} \tag{16.37}$$

Equation (16.37) not only solves the problem stated, but also makes it possible to find on the basis of Eq. (16.36) the minimum error variance:

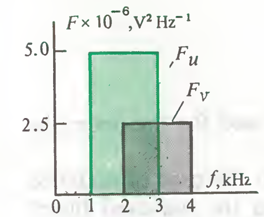
$$\sigma_{e,\,\mathrm{min}}^2 = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{W_x W_z}{W_x + W_z}\, d\omega \tag{16.38}$$

or, on passing from W_x and W_z to the one-sided power spectra F_x and F_z,

$$\sigma_{e,\,\mathrm{min}}^2 = \int_0^{\infty} \frac{F_x F_z}{F_x + F_z}\, df \tag{16.39}$$

Physical interpretation of the frequency properties of an optimum filter

The meaning of the above results is simple: The amplitude response of an optimum filter which minimizes the root-mean-square error must be large at the frequencies where the bulk of the signal energy is concentrated. Conversely, it must decrease where the power spectrum of the error is high.



Example 16.2. *A random process $X(t)$ (the wanted signal) has a band-limited one-sided power spectrum F_x equal to 5×10^{-6} $V^2 Hz^{-1}$ in the frequency band from 1 to 3 kHz and zero elsewhere. The random process $Z(t)$ (the noise) has a one-sided power spectrum showing a similar frequency dependence: $F_z = 2.5 \times 10^{-6}$ $V^2 Hz^{-1}$ in the frequency band from 2 to 4 kHz. Find the frequency response of the optimum filter and the minimum root-mean-square error in the signal reproduced.*

Taking advantage of Eq. (16.37), we find that the amplitude response of the optimum filter must be nonzero only in the frequency interval 1-3 kHz where the power spectrum of the detected signal is concentrated, such that

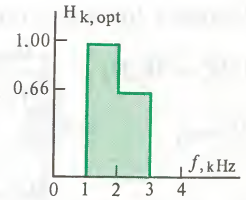
$$H_{k,\mathrm{opt}}(f) = \begin{cases} 1, & 1\ \mathrm{kHz} < f < 2\ \mathrm{kHz} \\ 0{,}66, & 2\ \mathrm{kHz} < f < 3\ \mathrm{kHz} \end{cases}$$

The signal variance equal to the product of the power spectrum by the frequency band occupied, is

$\sigma_x^2 = 5 \times 10^{-6} \times 2 \times 10^3 = 10^{-2}\ \mathrm{V}^2$

On the other hand, from Eq. (16.39) we get

$$\sigma_{e,\,\min}^2 = \frac{12.5 \times 10^{-12}}{7.5 \times 10^{-6}} \times 10^3 = 1.66 \times 10^{-3}\ \mathrm{V}^2$$



To sum up, when the two above random processes are subjected to linear filtering, the root-mean-square error in the signal reproduced will be at least 16.6%.

Summary

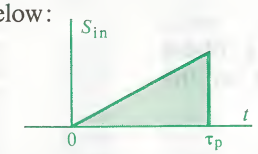
- A matched filter intended to detect a known signal mixed with Gaussian white noise at a maximum signal-to-noise ratio must have a frequency response proportional to the complex conjugate spectrum of the signal.
- The impulse response of a matched filter is a "mirror image" of the detected signal.
- The response of a matched filter to the signal being detected is proportional to the time-shifted autocorrelation function of the signal.
- The maximum attainable signal-to-noise ratio at the output of a linear filter depends only on the signal energy and the power spectrum of the noise.
- Satisfactory results in the detection of simple signals can be obtained by using quasi-optimum filters instead of matched filters.
- For a linear filter to detect a random signal with a minimum root-mean-square error, its amplitude response must be a maximum in the frequency band where the power spectrum of the signal is high, and it must be a minimum where the power spectrum of the noise is high.

Review Questions

1. What is the difference between the criterion of optimality for filters used for the detection of known and unknown signals?
2. Explain the principle of coherent addition of spectral components in the case of matched linear filtering. What is the difference between the comb filter and the matched filter?
3. State the condition for the delay time t_0 when a known signal is detected by an optimum filter.
4. What autocorrelation function of a signal should be like so that a matched filter could detect it effectively?
5. What principle underlies the construction of real matched filters for LFM pulses?
6. What is the relation of the minimum error variance to the power spectra of the signal and of the noise in the optimum filtering of a random signal?

Problems

1. Plot the impulse response of the filter matched to the triangular input signal shown below:



Give the minimum value of delay time t_0.

2. A resistor $R = 500\ \Omega$ held at $T = 300$ K acts as an equivalent source of white noise at the input of a filter. What should the resultant signal energy be for the matched filter to give a signal-to-noise ratio of $q = 5$?
3. A PSK Barker sequence consists of 13 pulses, each pulse being 10 μV in amplitude and 5 μs in duration. Find the power spectrum of the input white noise at which the output signal-to-noise ratio is unity.
4. There is an LFM pulse with a rectangular envelope, a frequency deviation of $\Delta\omega = 3 \times 10^7\ \mathrm{s}^{-1}$ and a duration of $\tau_p = 3.5 \times 10^{-5}$ s. Find the duration of the main lobe of the wave at the output of the filter matched to this signal.
5. The signal is a realization of a random process $X(t)$ for which the one-sided power spectrum is $F_x = 8 \times 10^{-5}\ \mathrm{V}^2\,\mathrm{Hz}^{-1}$ in the frequency band $0 < f < 300$ Hz, and zero elsewhere. This signal is combined with white noise, $Z(t)$, whose one-sided power spectrum is $F_z = 2 \times 10^{-6}\ \mathrm{V}^2\,\mathrm{Hz}^{-1}$ at all frequencies. Find the error variance in the signal detected by an optimum filter.

Advanced Problems

6. Draw a block diagram for the matched filter intended for the optimum detection of a triangular signal (see Problem 1).
7. Contemplate the possibility of reducing the side lobes of the autocorrelation function of an LFM signal by using time variations in frequency more complex than linear.
8. Derive an equation defining the frequency response of an optimum filter which can predict the instantaneous values of the signal $x(t + t_0)$, that is, extrapolate some "future" values of the signal in the best possible manner (in the mean-square sense).

17-728

Appendices

Appendix 1 Walsh Functions and Some of Their Properties

At present, several methods are in use for determining Walsh functions [22]. The most convenient way is one based on the recurrence relation

$$\mathrm{wal}(2n+p, \vartheta) = (-1)^{[n/2]+p}[\mathrm{wal}(n, 2\vartheta + 1/2)$$
$$+(-1)^{n+p}\,\mathrm{wal}(n, 2\vartheta - 1/2)],\ n = 0, 1, 2, \ldots \tag{A1.1}$$

Here $[n/2]$ stands for the largest integer number, which is smaller than or equal to $n/2$; the number p may take on values 0 or 1.
In carrying out iterations, it should be remembered that the function $\mathrm{wal}(0, \vartheta)$ is constant over the interval $-1/2 < \vartheta < 1/2$:

$$\mathrm{wal}(0, \vartheta) = \begin{cases} 0, & \vartheta < -1/2 \\ 1, & -1/2 < \vartheta < 1/2 \\ 0, & \vartheta > 1/2 \end{cases} \tag{A1.2}$$

For example, on setting $n = 0$ and $p = 1$, we obtain from (A1.1)

$$\mathrm{wal}(1, \vartheta) = -\mathrm{wal}(0, 2\vartheta + 1/2) - \mathrm{wal}(0, 2\vartheta - 1/2)$$

The Walsh functions have an interesting property:

$$\mathrm{wal}(m, \vartheta)\,\mathrm{wal}(n, \vartheta) = \mathrm{wal}(l, \vartheta) \tag{A1.3}$$

where the index l is the modulo 2 sum of the indexes m and n (symbolized as $l = m \oplus n$). In order to take a modulo 2 sum, the numbers m and n must be expressed in binary notation; then the binary numbers are added together without a carry to the next more significant bit position by the following rule:

$$1 \oplus 0 = 0 \oplus 1 = 1 \tag{A1.4}$$

$$0 \oplus 0 = 1 \oplus 1 = 0$$

For example, if $m = 5$ (dec.) = 101 (binary) and $n = 6$ (dec.) = 110 (binary), then $101 \oplus 110 = 011$ (binary) = 3 (dec.). Thus,

$$\mathrm{wal}(5, \vartheta)\,\mathrm{wal}(6, \vartheta) = \mathrm{wal}(3, \vartheta)$$

In the literature on the subject, in addition to the Walsh functions, use is often made of two more related systems, namely the even functions $\mathrm{cal}(n, \vartheta)$ (similar to cosines) and the odd fucntions $\mathrm{sal}(n, \vartheta)$ (similar to sines). The three classes of functions are related in the following manner:

$$\mathrm{cal}(n, \vartheta) = \mathrm{wal}(2n, \vartheta) \tag{A1.5}$$

$$\mathrm{sal}(n, \vartheta) = \mathrm{wal}(2n - 1, \vartheta)$$

Appendix 2 Berg Functions $\gamma_0(\vartheta)$, $\gamma_1(\vartheta)$, and $\gamma_2(\vartheta)$

$\vartheta°$	γ_0	γ_1	γ_2
0	0	0.000	0.000
10	0.000	0.001	0.001
20	0.004	0.009	0.008
30	0.015	0.029	0.027
40	0.034	0.066	0.056
50	0.065	0.121	0.095
60	0.109	0.196	0.138
70	0.166	0.287	0.176
80	0.236	0.390	0.203
90	0.318	0.500	0.212
100	0.410	0.611	0.203
110	0.509	0.713	0.176
120	0.609	0.805	0.138
130	0.708	0.878	0.095
140	0.800	0.934	0.056
150	0.881	0.971	0.027
160	0.944	0.991	0.008
170	0.985	0.999	0.001
180	1.000	1.000	0.000

Appendix 3 The Subroutine for the Computation of Berg Functions

```
SUBROUTINE BERG (N, T, G)
PI = 3.1415926
A = N
IF (N. EQ. 0) GOTO 1
IF (N. EQ. 1) GOTO 2
G = 2. * (SIN (A * T) * COS (T) − A * COS
     * (A * T) * SIN (T)/(PI * A * (A * A − 1.))
     RETURN
1    G = (SIN (T) − T * COS (T))/PI
     RETURN
2    G = (T − SIN (T) * COS (T))/PI
     RETURN
     END
```

This subroutine is called out with the CALL BERG (α, β, γ) statement. Here α and β are the actual values of the harmonic number n and the cutoff angle ϑ, respectively. The computed values of the function $\gamma_n(\vartheta)$ are stored in the "γ" memory location.

Appendix 4 Laplace Transform Pairs

$F(p)$	$f(t)$
1	$\delta(t)$
$1/p$	$\sigma(t)$
$1/p^2$	t
$1/(p+a)$	$\exp(-at)$
$p/(p+a)$	$\delta(t) - a\exp(-at)$
$a/[p(p+a)]$	$1-\exp(-at)$
$1/(p+a)(p+b)$	$\dfrac{1}{b-a}(e^{-at}-e^{-bt})$
$p/(p+a)(p+b)$	$\dfrac{1}{b-a}(be^{-bt}-ae^{-at})$
$1/(p+a)^2$	te^{-at}
$p/(p+a)^2$	$(1-at)e^{-at}$
$\omega/(p^2+\omega^2)$	$\sin\omega t$
$p/(p^2+\omega^2)$	$\cos\omega t$
$\omega/[(p+a)^2+\omega^2]$	$e^{-at}\sin\omega t$
$(p+a)/[(p+a)^2+\omega^2]$	$e^{-at}\cos\omega t$
$p/(p^2-a^2)$	$\cosh at$
$a^2/p^2(p+a)$	$at-(1-e^{-at})$
$\dfrac{1}{p(p+a)(p+b)}$	$\dfrac{1}{ab}\left[1+\dfrac{1}{a-b}(be^{-at}-ae^{-bt})\right]$
$\dfrac{1}{p[(p+a)^2+\omega^2]}$	$\dfrac{1}{a^2+\omega^2}[1-e^{-at}(\cos\omega t+\dfrac{a}{\omega}\sin\omega t)]$
$\dfrac{1}{(p+a)(p^2+\omega^2)}$	$\dfrac{1}{a^2+\omega^2}\left(e^{-at}-\cos\omega t+\dfrac{a}{\omega}\sin\omega t\right)$
$\dfrac{p}{(p+a)(p^2+\omega^2)}$	$\dfrac{1}{a^2+\omega^2}(-ae^{-at}+a\cos\omega t+\omega\sin\omega t)$
$\dfrac{p^2}{(p+a)(p^2+\omega^2)}$	$\dfrac{1}{a^2+\omega^2}(a^2e^{-at}-a\omega\sin\omega t+\omega^2\cos\omega t)$

Appendix 5 The Subroutine for the Computation of Discrete Fourier Coefficients

```
      SUBROUTINE TRFUR (X, A, B, N, G)
      DIMENSION X(N), A(N), B(N)
      G = 0.
      A1 = 0.
      AN = N
      DO1I = 1, N
      G = G + X(I)
1     A1 = A1 + (-1.)**I*X(I)
      G = G/AN
      A(N/2) = A1/AN
      M = N/2 - 1
      DO2I = 1, M
      AI = I
      A(I) = 0.
      B(I) = 0.
      DO3J = 1, N
      AJ = J
      T = 6.283185*AI*AJ/AN
      C = COS(T)
      S = SIN(T)
      A(I) = A(I) + X(J)*C
      A(I) = A(I) + X(J)*C
3     B(I) = B(I) + X(J)*S
      A(I) = A(I)*2./AN
2     B(I) = B(I)*2./AN
      RETURN
      END
```

The arguments of the above array are as follows: X, the real array of numbers being processed; A and B, the arrays of real and imaginary parts of the discrete Fourier transform, respectively; N, the length of the input array; G, the name of the variable corresponding to the constant component of the discrete Fourier transform. Some of the memory locations assigned to arrays A and B remain vacant, which is permitted if the length of the array being processed does not exceed several tens of samples.

Bibliography

1. GONOROVSKY I.S. *Radio Circuits and Signals.* Moscow, Mir Publishers, 1981 (English translation).
2. ZINOVYEV A.L. and FILIPPOV L.I. *An Introduction to Signal and Circuit Theory.* Moscow, Vysshaya Shkola, 1975 (in Russian).
3. TIKHONOV V.I. *Statistical Communication Theory.* Moscow, Sovyetskoye Radio, 1966 (in Russian).
4. KOTELNIKOV V.A. *The Theory of Optimum Noise Immunity.* New York, McGraw Hill Book Co., 1959 (in English).
5. WOZENCRAFT J.M. and JACOBS I.M. *Principles of Communication Engineering.* New York, London, John Wiley & Sons, 1965 (in English).
6. KOLMOGOROV A.N. and FOMIN S.V. *Elements of the Theory of Functions and Functional Analysis.* Moscow, Nauka, 1972 (in Russian).
7. MANDELSHTAM L.I. *Lectures on the Theory of Oscillations.* Moscow, Nauka, 1972 (in Russian).
8. STEPANOV V.V. *A Course in Differential Equations.* Moscow, GITTL, 1953 (in Russian).
9. TIKHONOV A.N., VASILYEVA A.B., and SVESHNIKOV A.G. *Differential Equations.* Moscow, Nauka, 1980 (in Russian).
10. GAKHOV F.D. *Boundary-Value Problems.* Moscow, Fizmatguiz, 1963 (in Russian).
11. LAVRENTYEV M.A. and SHABAT B.V. *Methods of the Theory of Functions of a Complex Variable.* Moscow, Fizmatguiz, 1958 (in Russian).
12. BOGOLIUBOV N.N. and MITROPOLSKY Yu.A. *Asymptotic Methods in the Theory of Nonlinear Oscillations.* Moscow, Fizmatguiz, 1958 (in Russian).
13. KOBZAREV Yu.B. *On a Nonlinear Treatment of Events in a Valve Oscillator.* ZhTF, No. 5 (1933) (in Russian).
14. RYTOV S.M. *An Introduction to Statistical Radio Physics. Part I.* Moscow, Nauka, 1976 (in Russian).
15. LEVIN B.R. *Theoretical Foundations of Statistical Communication. Book 1,* Moscow, Sovyetskoye Radio, 1974 (in Russian).
16. FRANKS L.E. *Signal Theory.* Englewood Cliffs, New Jersey, Prentice-Hall Inc., 1969.
17. PAPOULIS A. *Systems and Transforms with Applications in Optics.* New York, McGraw Hill Book Co., 1967.
18. GOLDMAN S. *Information Theory.* London, Constable and Co., 1953.
19. BENNET W.R. and DAVEY J.R. *Data Transmission.* New York, McGraw Hill Book Co., 1965.
20. MIDDLETON D. *An Introduction to Statistical Communication Theory.* New York, McGraw Hill Book Co., 1960.
21. COOK C.E. and BERNFELD M. *Radar Signals.* New York, Academic Press, 1967.
22. HARMUTH H.R. *Transmission of Information by Orthogonal Functions.* Berlin, Heidelberg, New York, 1970.
23. KHARKEVICH A.A. *Interference Control.* Moscow, Nauka, 1965 (in Russian).
24. VARAKIN L.Ye. *A Theory of Complex Signals.* Moscow, Sovyetskoye Radio, 1970 (in Russian).

25. ZERNOV N.V. and KARPOV V.G. *Circuit Theory.* Moscow, Energiya, 1965 (in Russian).
26. BASKAKOV S.I. *Distributed-Parameter Circuits.* Moscow, Vysshaya Shkola, 1980 (in Russian).
27. GUILLEMIN E.A. *Synthesis of Passive Networks.* New York, John Wiley & Sons, Inc., 1960.
28. KARNI S. *Networks Theory: Analysis and Synthesis.* Boston, Massachusetts, Allun and Bacon, Inc., 1966.
29. WAI-KAI CHEN. *Theory and Design of Broadband Matching Networks.* Oxford, Pergamon Press, 1976.
30. MATKHANOV P.N. *Synthesis of Linear Electric Circuits.* Moscow, Vysshaya Shkola, 1976 (in Russian).
31. ANDREYEV V.S. *Theory of Nonlinear Electric Networks.* Moscow, Svyaz, 1972 (in Russian).
32. MASLENNIKOV V.V. and SIROTKIN A.P. *Selective RC Amplifiers.* Moscow, Energiya, 1980 (in Russian).
33. HUELSMAN L.P., ed. *Active Filters: Lumped, Distributed, Integrated, Digital and Parametric.* New York, McGraw-Hill Book Co., 1970.
34. PELED A. and LIU B. *Digital Signal Processing.* New York, John Wiley & Sons, Inc., 1976.
35. BOGNER R.E. and CONSTANTINIDES A.G., eds. *Introduction to Digital Theory.* London, John Wiley & Sons, Inc., 1976.
36. BODE H.W. *Network Analysis and Feedback Amplifiers Design.* New York, D. van Nostrand Co., 1945.
37. KOTELNIKOV V.A., and NIKOLAYEV A.M. *Foundations of Communication Engineering. Part II,* Moscow, Svyazizdat, 1954 (in Russian).
38. TIKHONOV V.I., ed. *Statistical Communication. Examples and Problems.* Moscow, Sovietskoye Radio, 1980 (in Russian).
39. NIKOLAYEV A.M., ed. *Problems in Signals and Circuits.* Moscow, Sovyetskoye Radio, 1972 (in Russian).
40. GRADSHTEIN I.S. and RYZHIK I.M. *Tables of Integrals, Sums, Series and Products.* Moscow, GIFML, 1963 (in Russian).
41. LOSEV A.K. *An Introduction to Communication Engineering.* Moscow, Vysshaya Shkola, 1980 (in Russian).
42. PONTRYAGUIN L.S. *A Mathematical Theory of Optimum Processes.* Moscow, Nauka, 1976 (in Russian).
43. STRUTT J.W. (Lord Rayleigh). *Theory of Sound.*

Index